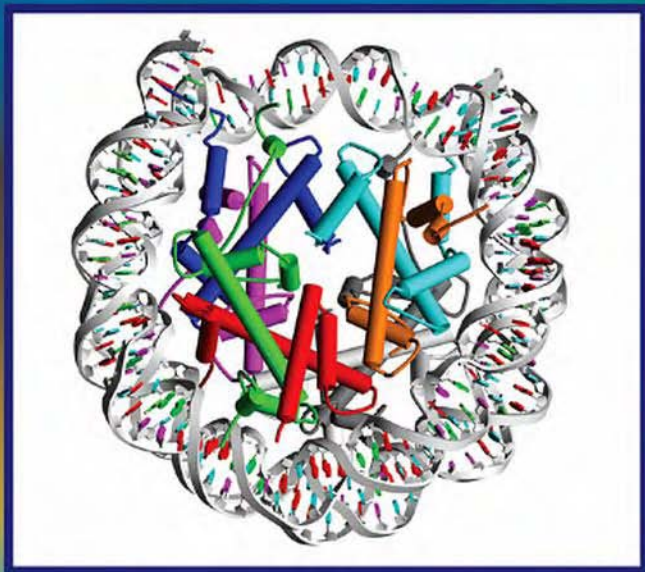


Principles of Nucleic Acid Structure



Stephen Neidle



VISIT...

LANZAROTE
Caliente.COM

Principles of Nucleic Acid Structure

This page intentionally left blank

Principles of Nucleic Acid Structure

Stephen Neidle

The School of Pharmacy
University of London, London, UK



ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Academic Press is an imprint of Elsevier



Academic Press is an imprint of Elsevier
84 Theobald's Road, London WC1X 8RR, UK
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
Linacre House, Jordan Hill, Oxford OX2 8DP, UK
30 Corporate Drive, Suite 400, Burlington, MA 01803, USA
525 B Street, Suite 1900, San Diego, CA 92101-4495, USA

First edition 2008

Copyright © 2008 Elsevier Inc. All rights reserved

Material in this book originally published in "*Nucleic Acid Structure and Recognition*", by Stephen Neidle (Oxford University Press, 2002)

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means electronic, mechanical, photocopying, recording or otherwise without the prior written permission of the publisher

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone (+44) (0) 1865 843830; fax (+44) (0) 1865 853333; email: permissions@elsevier.com. Alternatively you can submit your request online by visiting the Elsevier web site at <http://elsevier.com/locate/permissions>, and selecting *Obtaining permission to use Elsevier material*

Notice

No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein. Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

CIP applied for and in process

ISBN: 978-0-12-369507-9

For information on all Academic Press publications visit our website at books.elsevier.com

Printed and bound in USA

08 09 10 11 12 10 9 8 7 6 5 4 3 2 1

On the Cover:

Structure of the nucleosome core particle, drawn from coordinates taken from PDB entry no. 1KX3 (Davey et al., Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. J. Mol. Biol. 2002, 319, 1097–1113).

Working together to grow
libraries in developing countries

www.elsevier.com | www.bookaid.org | www.sabre.org

ELSEVIER

BOOK AID
International

Sabre Foundation

*To the memory of my father,
who inspired my curiosity for science*

This page intentionally left blank

Preface

The years that have elapsed since the previous version of this book was published, in 2001, have been momentous ones for nucleic acid studies. In 2003 we celebrated both the 50th anniversary of the discovery of the structure of the DNA double helix, and the announcement of the determination of the sequence of the human genome. It might therefore be thought that the study of nucleic acid structure is itself now part of history, and that there is little more to be known. The reality is very different; we have seen a number of profound new discoveries relating to both RNA and DNA structure, just in the first seven years of this millennium. These significant advances in the subject have required, not just a new edition, but an expansion of many sections and a re write of others.

The aim of the book is to provide an introduction to the underlying fundamental features and principles governing nucleic acid structures, as well as many of the structures themselves. It is hoped that this provides a firm foundation for subsequent studies of the structural biology and chemistry of nucleic acids. Its intended audience is at graduate level, and it is hoped that it will be of use to active researchers, and even to the more inquisitive final-year undergraduate students. The book does not attempt to be a comprehensive survey of all nucleic acid-containing structures. Instead, it concentrates on more general themes, and focuses on those structures that illustrate a particular feature of interest or generality, especially in the context of their relevance to chemical, biological, or pharmacological issues. I apologize in advance to those whose favourite structure has been ignored in favour of my own more subjective judgments.

The book emphasizes those structures determined by X-ray crystallography, since this methodology continues to dominate the field in terms of size of molecule whose structure can be determined, as well as still providing the majority of high-resolution structures. The introduction to crystallography and other techniques is designed to provide the non-specialist with sufficient understanding to read the primary literature, and most importantly, to be able to begin to judge the scope and quality of both experimental and theoretical structural studies. I have also expanded the reference and reading lists to provide a reasonably comprehensive guide to both the past and recent literature, and have included information on a number of relevant websites.

Any book on molecular structure suffers from the disadvantage of not being able to adequately convey the three-dimensionality of structures. The previous edition was associated with a dedicated Internet site, which enabled the structures to be examined interactively, and in a variety of display modes. The excellence of the many graphics programs freely available on the web, together with the molecular display tools available from the Protein Data Bank and other web sites, makes a dedicated site no longer

necessary, or even desirable. I have included tables of PDB and NDB (Nucleic Acid Database) codes for a large number of representative structures, to aid the reader in speedily viewing a particular feature, or downloading a structure file for subsequent display and analysis on one's own desktop or laptop. I have also included a list of my own favourite molecular graphics programs that have nucleic acid-friendly features.

I am grateful to my wife Andrea and children Dan, Ben, and Hannah for their constant support and encouragement in this and many other ventures, and to my colleagues, collaborators, and students for their contributions, insights, and discussions. Thanks also to my editor at Elsevier, Kirsten Funk, for all her hard work, patience and support.

Stephen Neidle
London, June 2007

Contents

1. Methods for Studying Nucleic Acid Structure	1
1.1 Introduction	1
1.2 X-ray Diffraction Methods for Structural Analysis	2
1.2.1 Overview	2
1.2.2 Fiber Diffraction Methods	5
1.2.3 Single-Crystal Methods	7
1.3 NMR Methods for Studying Nucleic Acid Structure and Dynamics	10
1.4 Molecular Modelling and Simulation of Nucleic Acids	11
1.5 Chemical, Enzymatic, and Biophysical Probes of Structure and Dynamics	14
1.6 Sources of Structural Data	15
1.7 Visualization of Nucleic Acid Molecular Structures	15
1.7.1 The Structures in This Book	16
2. The Building-Blocks of DNA and RNA	20
2.1 Introduction	20
2.2 Base Pairing	23
2.3 Base and Base Pair Flexibility	24
2.4 Sugar Puckers	28
2.5 Conformations About the Glycosidic Bond	32
2.6 The Backbone Torsion Angles and Correlated Flexibility	33
3. DNA Structure as Observed in Fibers and Crystals	38
3.1 Structural Fundamentals	38
3.1.1 Helical Parameters	38
3.1.2 Base-Pair Morphological Features	38
3.2 Polynucleotide Structures from Fiber Diffraction Studies	39

3.2.1 Classic DNA Structures	39
3.2.2 DNA Polymorphism in Fibers	43
3.3 B-DNA Oligonucleotide Structure as Seen in Crystallographic Analyses	47
3.3.1 The Dickerson–Drew Dodecamer	47
3.3.2 Other Studies of the Dickerson–Drew Dodecamer	49
3.3.3 Other B-DNA Oligonucleotide Structures	51
3.3.4 Sequence-Dependent Features of B-DNA: Their Occurrence and Their Prediction	56
3.4 A-DNA Oligonucleotide Crystal Structures	60
3.4.1 A-Form Octanucleotides	60
3.4.2 Do A-Form Oligonucleotides Occur in Solution? Crystal-Packing Effects	61
3.4.3 The A $\leftrightarrow$ B Transition in Crystals	63
3.5 Z-DNA – Left-Handed DNA	64
3.5.1 The Z-DNA Hexanucleotide Crystal Structure	64
3.5.2 Overall Structural Features	65
3.5.3 The Z-DNA Helix	66
3.5.4 Other Z-DNA Structures	67
3.5.5 Biological Aspects of Z-DNA	67
3.6 Bent DNA	69
3.6.1 DNA Periodicity in Solution	69
3.6.2 A-Tracts and Bending	70
3.6.3 Structures Showing Bending	71
3.6.4 The Structure of Poly dA•dT	73
3.7 Concluding Remarks	74
4. Nonstandard and Higher-Order DNA Structures: DNA–DNA Recognition	81
4.1 Mismatches in DNA	81
4.1.1 General Features	81
4.1.2 Purine:Purine Mismatches	82
4.1.3 Alkylation Mismatches	85
4.2 DNA Triple Helices	88
4.2.1 Introduction	88
4.2.2 Structural Studies	90
4.2.3 Antiparallel Triplexes and Nonstandard Base-pairings	95
4.2.4 Triplex Applications	100
4.3 Guanine Quadruplexes	101
4.3.1 Introduction	101
4.3.2 Overall Structural Features of Quadruplex DNA	103
4.3.3 Examples of Simple Quadruplex Structures	107

4.3.4	Some Complex Quadruplex Structures	108
4.3.5	The i-Motif	113
4.4	DNA Junctions	114
4.4.1	Holliday Junction Structures	114
4.4.2	DNA Enzyme Structures	118
4.5	Unnatural DNA Structures	120
5.	Principles of Small Molecule-DNA Recognition	132
5.1	Introduction	132
5.2	DNA-Water Interactions	136
5.2.1	Hydration in the Grooves in Detail	140
5.3	General Features of DNA-Drug and Small-Molecule Recognition	143
5.4	Intercalative Binding	144
5.4.1	Simple Intercalators	146
5.4.2	Complex Intercalators	147
5.4.3	Major-Groove Intercalation	151
5.4.4	Bis-Intercalators	158
5.5	Intercalative-Type Binding to Higher-Order DNAs	163
5.5.1	Triplex DNA-Ligand Interactions	163
5.5.2	Ligand Binding to Quadruplex DNAs	164
5.5.3	Ligand Binding to Junction DNAs	166
5.6	Groove-Binding Molecules	169
5.6.1	Simple Groove Binding Molecules	169
5.6.2	Netropsin and Distamycin	178
5.6.3	Sequence-Specific Polyamides	182
5.7	Small Molecule Covalent Bonding to DNA	187
5.7.1	The Platinum Drugs	188
5.7.2	Covalent-Binding Combined with Sequence-Specific Recognition	191
6.	RNA Structures and Their Diversity	204
6.1	Introduction	204
6.2	Fundamentals of RNA Structure	206
6.2.1	Helical RNA Conformations	206
6.2.2	Mismatched and Bulged RNA Structures	210
6.3	Transfer RNA Structures	217
6.4	Ribozymes	221
6.4.1	The Hammerhead Ribozyme	223
6.4.2	Complex Ribozymes	224
6.5	Riboswitches	227

6.6 The Ribosome, a Ribozyme Machine	229
6.6.1 The Structure of the 30S Subunit	232
6.6.2 The Structure of the 50S subunit	234
6.6.3 Complete Ribosome Structures	234
6.7 RNA-Drug Complexes	235
6.8 RNA Motifs	241
7. Principles of Protein-DNA Recognition	249
7.1 Introduction	249
7.2 Direct Protein-DNA Contacts	252
7.3 Major-Groove Interactions – the α -Helix as the Recognition Element	257
7.4 Zinc-Finger Recognition Modes	259
7.5 Other Major Groove Recognition Motifs	263
7.6 Minor-Groove Recognition	264
7.6.1 Recognition of B-DNA	264
7.6.2 The Opening-up of the Minor Groove by TBP	267
7.6.3 Other Proteins that Induce Bending of DNA	268
7.7 DNA-Bending and Protein Recognition	272
7.8 Protein-DNA-Small Molecule Recognition	275
Index	283

Methods for Studying Nucleic Acid Structure

1.1 Introduction

Our knowledge of DNA and RNA three-dimensional structure has advanced immeasurably since the elucidation of the first such structure, that of the DNA double helix in 1953 by Watson and Crick in conjunction with the X-ray fiber diffraction data of Franklin and Wilkins. Fiber diffraction methods subsequently enabled the morphologies of a whole range of nucleic acid double helical types to become established. More recently, the relationships between DNA primary sequence and the fine details of its molecular structure have become increasingly understood, in large part from single-crystal and nuclear magnetic resonance (NMR) structural studies on defined-sequence oligonucleotides. DNA structure continues to surprise with its ability to exist in a wide variety of forms, such as left-handed and multiple-stranded helices. The study of RNA structure has a more recent history, which has revealed that RNA can fold in a wide variety of complex ways as well as occur in double-helical form. There is now a very large amount of experimental information on the structures of protein-DNA, protein-RNA, and drug-DNA complexes.

The discovery of the double helix, as Watson and Crick realized, immediately provided fundamental new insights into the nature of genetic events. We now have extensive knowledge of both the detail and the variety of DNA and RNA structures themselves, together with the manner in which they are recognized by regulatory, repair, and other proteins, as well as by small molecules. All this is giving us altogether more profound levels of understanding of the processes of gene regulation, transcription and translation, mutation/carcinogenesis, and drug action at the atomic and molecular levels. We are now beginning to piece together how all this works in the context of eukaryotic chromatin, so the challenges over the next few years will be to study the structural biology of large-scale DNA-protein structural assemblies just as has already been done for the ribosome.

These advances in nucleic acid structural studies have been largely due to the increased power and sophistication of the experimental approach of X-ray crystallography, which have provided most of the highly detailed structural information to date. The dominance of the crystallographic approach still continues, and is reflected in the emphasis of this book. NMR spectroscopy, molecular modeling/simulation,

and chemical/biochemical probe techniques also play important roles in providing information on structure, dynamics, and flexibility that can approach near-atomic resolution in at least some of its detail. Traditional spectroscopic-based biophysical methods can provide important complementary information, mostly at the macroscopic level. More recently developed techniques such as surface plasmon resonance spectroscopy and single-molecule methods, are extending their power so that the gap is now diminishing between macroscopic data on nucleic acids which the more traditional methods provide, and that at the atomic level. Underlying all of this progress have been the significant technical advances, notably in (i) the development of routine chemical methods for oligonucleotide synthesis and purification at the milligram level for both DNA and the more demanding RNA sequences, and (ii) the advent of efficient (and increasingly routine) cloning and expression systems for RNA and DNA-binding proteins, and for native RNA molecules.

This chapter provides a brief introduction to the two major structural methods, emphasizing their scope as well as their limitations for nucleic acid structural studies.

1.2 X-ray Diffraction Methods for Structural Analysis

1.2.1 Overview

X-rays typically have a wavelength of the same dimensions as interatomic bonds in molecules (about 1.5 Å). Scattering (or diffraction) of X-rays by molecules in ordered matter is the result of interactions between the radiation and the electron distribution of each component atom. Typical **diffraction patterns** from DNA, in the form of fibers or single crystals, are shown in Figs. 1.1, and 1.2. Reconstruction of the internal molecular arrangement by analysis of the scattered X-rays, analogous to a lens focusing scattered light from a microscope sample, thus provides a picture of the **electron density** distribution in the molecule. This reconstruction is not generally straightforward, due to the loss of phase information from the individual reflected X-rays during the diffraction process. The **phase problem** needs to be solved (see later for a brief description of various methods of doing this) in order for the electron density to be calculated in three dimensions (as a Fourier series), which is commonly termed a **Fourier map**. The approximate equivalence of the wavelength of X-rays and bond distance, of ~1–1.5 Å, means that in principle, the electron density of individual atoms in a molecule can be resolved, provided that the pattern of diffracted X-rays can be reconstituted into a real-space image.

The degree of electron density detail that can actually be seen is dependent on the **resolution** of the recorded diffraction pattern. Resolution may be defined in terms of the shortest separation between objects (i.e. atoms or groups of atoms in a molecule) that can be observed in the electron density reconstituted from the diffraction pattern. The resolution limit (r) is governed by the maximum diffraction angle (θ) recorded for the diffraction data and the wavelength (λ) of the X-rays: r is defined as $\lambda/2\sin\theta$. At a resolution of 2.5 Å, individual atoms in a structure cannot be resolved in an electron-density map although the shape and orientation of ring systems (e.g. base pairs) can be readily distinguished. These appear as elongated regions of electron density, with substituents being apparent as “outgrowths” from the main density. At 1.5 Å, individual atoms are generally just about observable in a map, although only at about 1.0–1.2 Å are all atoms fully

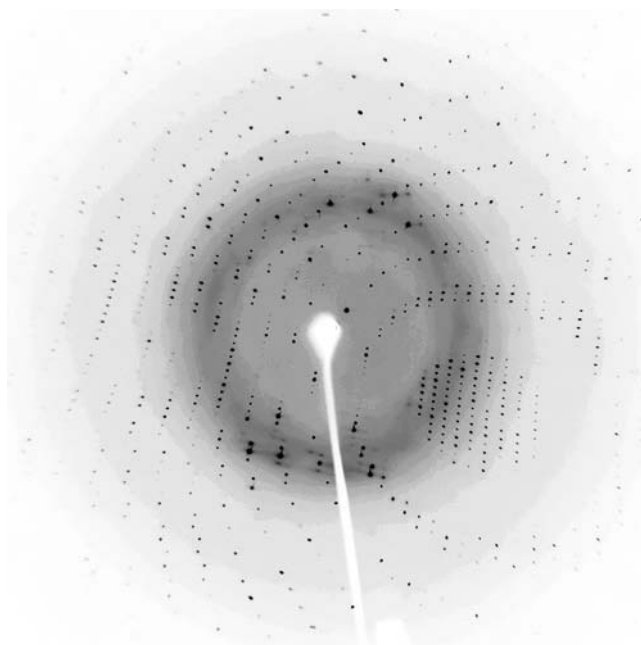


Figure 1.1 X-ray diffraction pattern from a single crystal of a drug-oligonucleotide complex, taken with an image plate and a conventional laboratory X-ray source. The resolution limit for this pattern is 1.6 Å.

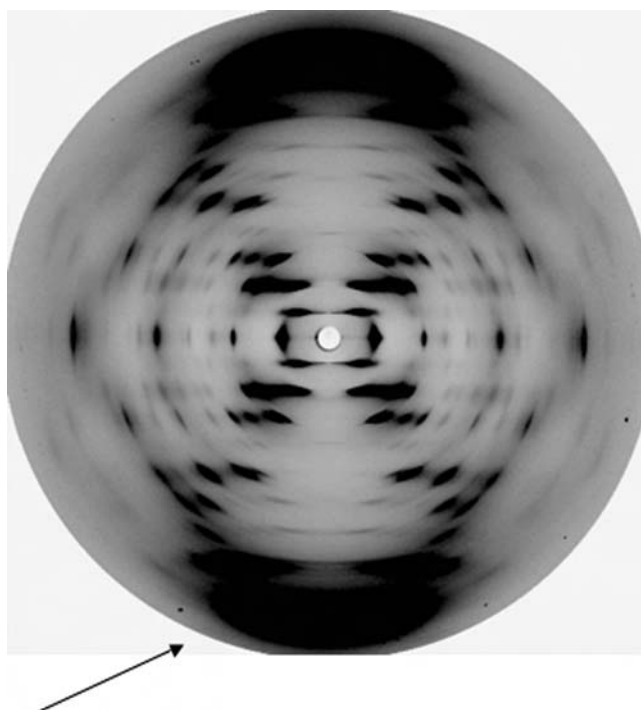


Figure 1.2 X-ray fiber-diffraction pattern from a sample of calf thymus DNA, showing a characteristic B-form pattern. (Provided by Professor Watson Fuller.) The arrow indicates the 3.4 Å layer line.

resolved and separated from each other (Fig. 1.3). There has been a marked increase in high-resolution studies (up to $0.8\text{\AA}$) in recent years due to the increasing use and world-wide availability of high-flux **synchrotron** sources of X-rays for structural biology studies. There are currently about 20 synchrotrons with beam lines dedicated to crystallography, with several more scheduled over the next few years (see <http://biosyn.rcsb.org> for further details). In 2005, of the 4515 macromolecular crystal structures submitted to the Protein Data Bank (<http://www.rcsb.org>), data for 3398 (75.3%) were collected on synchrotron beam lines. The proportion in 2006 is even higher (78.1%). Synchrotron beam lines have intensities of X-ray beams greater by several orders of magnitude than conventional laboratory X-ray sources. Synchrotron facilities have also enabled much smaller crystals than previously to be successfully analyzed. Most importantly, the ability to tune X-rays to differing wavelengths has provided the means whereby powerful methods of structure analysis can be employed (see later). Although not comparable with synchrotron beam intensities, the development of highly effective mirror-optics focusing systems for laboratory X-ray sources in recent years can enable diffraction data to be collected in-house, especially when larger crystals can be obtained. It is now almost universal practice to collect diffraction data from

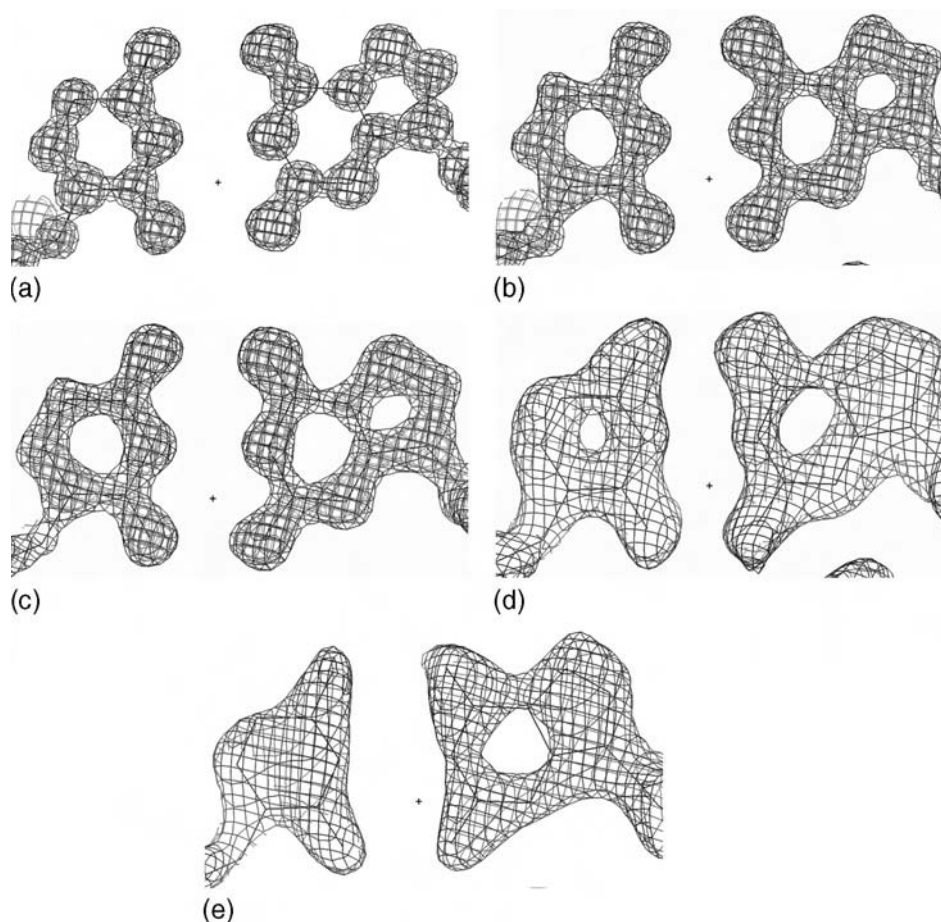


Figure 1.3 Calculated electron density in the plane of a C•G base pair, calculated at differing resolutions, showing the amount of atomic detail visible at particular resolutions: (a) $0.9\text{\AA}$, (b) $1.25\text{\AA}$, (c) $1.5\text{\AA}$, (d) $2.0\text{\AA}$ and (e) $2.5\text{\AA}$.

macromolecules at liquid nitrogen temperatures using the technique of **flash-freezing**, which tends to minimize crystal decay in the X-ray beam, and can improve the diffracting power of a given crystal.

The number of individual diffraction maxima observed from a crystalline or semicrystalline sample depends directly on two factors: the resolution of the pattern, and the size of the crystallographic **unit cell**. The number of unique reflections derived from these measurements also depends on the **symmetry** of the crystal. An ultra-high-resolution (0.74 Å) structure of a typical DNA 10-mer oligonucleotide crystal would give approximately 29,000 individual unique maxima (reflections) in a monoclinic space group. By contrast, a 12-mer oligonucleotide crystal in the well-studied orthorhombic space group $P2_12_12_1$, would give only some 3,000 unique reflections at 2.2 Å resolution, which has historically been a common limit for oligonucleotide crystals with diffraction data collected using laboratory X-ray sources. The measured intensity of an individual reflection is proportional to its **structure amplitude**, or observed **structure factor**, which when combined with phase information, results in the calculated structure factor. X-ray structures are always optimized (**refined**) against this observed structure amplitude data, usually by a least-squares method (see later).

The accuracy and reliability of the resulting structure depends in part on the quantity and resolution of the diffraction data, as well as the quality of its measurement. Of key importance is the actual correctness of the structural model itself, both in gross outline (which is defined by the low- to medium-resolution data), and in the detailed aspects of the structure (defined by the high-resolution data). The standard ways of assessing these factors for a given structural model are:

- To calculate the crystallographic **R factor**, defined as: $R = \Sigma ||F_o| - |F_c|| / \Sigma |F_o|$, summed over all observed reflections, where F_o and F_c are the observed and calculated structure factors. R , which is also termed the reliability index, is expressed as a percentage, or sometimes as a decimal.
- To calculate the so-called **free R factor** (R_{free}) for a small (typically 5%) subset of reflections, often chosen randomly. This set is not used in the refinement, and so the value of R_{free} is unbiased by the course of the refinement and any errors introduced during it.

The R value for a correctly refined structural model can range from <10% to >20%; in general the lower the value the more reliable is the model. Values for R_{free} are usually a few percent higher, but are very sensitive to even small changes and errors in the model. R_{free} is often used to judge the completion of a structure analysis, especially in terms of the behavior of water molecules located in electron-density maps and added to the model during successive rounds of refinement. At a certain point, adding more water molecules may well reduce the value of R simply because the number of variables is increased in the least-squares refinement: however if the value of R_{free} increases, then these “water molecules” are not physically real. It is also common practice to calculate Fourier maps with parts of a structure omitted to verify that these parts reappear in the map at the stereochemically correct positions.

1.2.2 Fiber Diffraction Methods

Historically, helical DNA and RNA structures were first analyzed by fiber diffraction techniques. Polymeric nucleic acids directly extracted from cell nuclei, have not been crystallized as single crystals capable of three-dimensional structure analysis. Initial

diffraction patterns of poor quality were obtained from DNA by Astbury in 1937/38, but significant progress was not made until the early 1950s, by Wilkins, Franklin, and their associates. They made DNA into oriented fibers, when the act of “pulling” such a fiber orients the nucleic acid helix along the direction of the fiber. These fibers can have exactly repetitious helical dimensions even though the underlying naturally occurring (often genomic) nucleic acid sequences in them are not simple repeats, and thus the sequence information in the nucleic acid molecules is lost. An important exception occurs with the use of synthetic polynucleotides of known, simply repeating sequence such as poly(dA-dT)•poly(dA-dT).

Natural and synthetic polynucleotides can form fibers with varying degrees of internal order, having one- or two-dimensional paracrystalline arrays in the fiber, with the latter usually having the greater order because of their nonrandom sequences. These differing degrees of order are reflected in their X-ray diffraction patterns, with natural double-helical DNA and RNA molecules usually having a degree of order along the **helix axis** but being randomly oriented with respect to each other. This gives rise to an X-ray diffraction pattern with characteristic spots and streaks of intensity, for example, the “helical cross” diffraction pattern, which is characteristic of B-type DNA double helices. Such patterns can be analyzed to give the **helical dimensions** of pitch, rise, and number of residues per helical turn, as well as defining the overall helical type (A, B, etc.). Even the best-ordered of paracrystalline polynucleotide fibers give at most only a few hundred individual diffraction maxima, corresponding to a typical maximum resolution of about 2.5 Å.

It is not in general possible to analyze this pattern of fiber diffraction intensities, determine phases, and derive a molecular structure *ab initio*, since the pattern is an average from all of the nucleotide units in a helical repeat. Instead, the pattern is fitted to a model using a least-squares procedure (Arnott, 1970). This enables conformational details of the averaged (mono- or dinucleotide) repeat to be varied and optimized. The correctness and quality of the model may be assessed using the standard crystallographic R factor. Values for R of 0.15–0.25 indicate that the calculated diffraction pattern agrees well with the observed one, and that the model is physically reasonable in terms of its stereochemistry.

An important question is whether the attainment of good agreement for these criteria necessarily means that the phase problem for these fiber structures has been uniquely solved. The process of analysis assumes a particular starting model and other models might in principle fit a data set at least moderately well, especially since atomic levels of resolution are not available from fiber diffraction. This question led some years ago to several suggestions of alternative structures to the Watson-Crick antiparallel double helix for DNA. Only on detailed examination was it found that none of these alternatives could be fitted in an acceptable manner to the observed diffraction data, as defined by the R factor and other tests. This, together with their numerous close intramolecular contacts, enabled these alternative structures to be conclusively rejected and the antiparallel double helix accepted as the sole model that can acceptably fit the B-DNA observed diffraction data.

It is striking, in spite of the limitations of fiber diffraction methods, that their characterizations of idealized DNA and RNA double-helical geometry in terms of helical type (A, B, etc.), have been found to be closely in accord with the large number of more recent single-crystal analyses of short helical segments, even at much higher resolutions, as well as of DNA/RNA helices in multi-subunit protein complexes such as the ribosome and nucleosomes. Fiber diffraction analysis also has the considerable

advantage of being able to readily study conformational transitions under a range of environmental conditions. The $A \leftrightarrow B$ transition of duplex DNA observed with variations in relative humidity is the classic example of this technique.

1.2.3 Single-Crystal Methods

By contrast with fiber analyses, single-crystal X-ray crystallographic methods are able to determine the complete three-dimensional molecular structures of biological macromolecules without necessarily recourse to any preconceived model, provided the molecules are discrete and not the effectively infinite disordered polymers of nucleic acid fibers. Single crystals can be thus defined as ordered arrays of discrete and identical molecules in three dimensions.

Crystallization of many DNA and RNA oligonucleotides has historically been challenging, being in the past sometimes more dependent on chance than systematic scientific study. This has changed radically with the advent of manual (and increasingly) automated methods to rapidly and systematically screen a wide set of crystallizing conditions (Ducruix and Giegé, 1999). A number of commercial kits are now available with pre-prepared solutions of a wide range of concentrations and types of counter-ion, buffer, and precipitating agents, so that a large number of crystallizing trials can be set up with minimal effort. This approach is also useful for finding alternative crystal forms if initial trials produce crystals with poor diffraction or exceptionally large unit cells. The use of robotic crystallization methods is increasingly common. These enable rapid, large-scale screening of crystallization conditions to be undertaken, which is especially useful when dealing with “difficult” molecules such as large RNAs or protein-nucleic acid multi-subunit complexes. Many nucleic acid crystallographers have developed specialized sets of conditions for their own speciality, notable examples being in the RNA and ribozyme field (e.g., Ke and Doudna, 2004).

The range of resolution reported for single-crystal studies of oligonucleotides spans from 0.7 Å to 3.0 Å (Fig. 1.3). Thus, the highest-resolution oligonucleotide structures have true atomic resolution and accordingly are of corresponding accuracy (≤ 0.02 Å for distances and $\leq 0.2^\circ$ for angles) in respect of derived geometric parameters. A typical 2.5 Å resolution structure analysis, by contrast, would have distances reliable to about ± 0.3 Å and angles to about $\pm 5^\circ$. However it is necessary to use constraints to standard bond geometries during the crystallographic refinement process of nonatomic resolution crystal structures. This means that it is not only nonbonded and intermolecular distances but also conformational and base morphological features that have to be interpreted with care, and likely errors and uncertainties taken into account. Hydrogen atoms are only directly observed in electron-density maps from the very highest-resolution oligonucleotide analyses, and so hydrogen-bonding schemes (especially those involving water molecules) normally have to be inferred.

X-ray diffraction patterns from oligonucleotide crystals can be analyzed, and their underlying molecular structures solved *ab initio* by the standard heavy-atom multiple and single **isomorphous replacement** (MIR and SIR) phasing methods of macromolecular crystallography. These do not presume any particular structural model and hence do not bias the resulting structure, for example, to, have all Watson-Crick base-pairing in a double-stranded oligonucleotide. However a number of such heavy-atom derivatives are required for satisfactory MIR phasing, which are not always readily obtained, especially for helical nucleic acids. In favorable cases it is possible to

solve a structure with a single derivative by means of a combination of phasing from isomorphous replacement and anomalous scattering at a single wavelength.

The availability of tunable-wavelength X-ray facilities at many high-flux synchrotron facilities has enabled the technique of phasing by **multi-wavelength anomalous diffraction** (MAD) to be used. This uses a single appropriate heavy atom, which has the ability to absorb X-rays to differing extents at different wavelengths; phases and hence electron-density maps can be directly calculated from such data. These maps, when obtained at high resolution, are sometimes of remarkably high quality, revealing complete structures at the outset. It is fortunate for nucleic acid crystallography that bromine and iodine atoms, which can be readily chemically attached to uracil bases, provide excellent anomalous diffraction signals. This powerful method is now the method of choice, limited only by the availability of sufficient tunable synchrotron beam time. Several alternatives to the use of bromine or iodine have been found, which are occasionally needed when it is found that the halogen–uracil bond is rapidly cleaved in the X-ray beam. Examples include using a single nucleotide with a thio-containing backbone (to bind a mercury heavy-atom derivative), or a phosphoselenoate to replace a phosphate (Wilds *et al.*, 2002). It is possible to replace oxygen by selenium at the 2' position of a thymine or uridine nucleotide (Jiang *et al.*, 2007; Pallan and Egli, 2007a), in the nonbridging backbone (Pallan and Egli, 2007b), or at the thymidine 4-position (Salon *et al.*, 2007). Oligonucleotides incorporating this selenium modification produce crystals that grow much more rapidly and having higher diffraction quality than do bromine-derivatized oligonucleotides. It is possible to utilize the anomalous signal of phosphorus atoms to phase nucleic acid structures when ultra-high-resolution ($<1.0\text{ \AA}$) diffraction data of high redundancy is available (Dauter and Adams, 2001).

Alternatively it is possible to take account of the fact that many nucleic acid structures crystallize in an arrangement isomorphous to structures previously determined (e.g., by heavy-atom phasing) or are presumed to contain a particular structural motif such as a double helix. These structures can often be solved by molecular replacement or “search” methods, which assume at least part of the structure and attempt to locate it in the crystallographic unit cell. Problems have occasionally arisen with this approach, when, for example, a helix has been correctly oriented within the unit cell but its position is incorrectly indicated, being systematically related to the correct one, for example by a simple translation of a base-pair. Search methods become increasingly challenging with a decreasing fraction of known geometry in a structure, and heavy-atom methods then become advisable. They are also difficult when the correct geometry of the search fragment is not precisely known, and then the correct rotational and translational solution becomes unclear. New protein–nucleic acid crystal structures are usually solved by heavy-atom, MIR or MAD methods, as are an increasing number of new types of oligonucleotide crystal structure. It is fortunate that the key crystal structures of a B and a Z-DNA oligonucleotide have been solved *ab initio* by heavy-atom methods (see Chap. 3), thereby ensuring a firm and unambiguous basis for subsequent molecular replacement analyses of a large number of DNA oligonucleotide structures.

The increasing possibility of obtaining true atomic-resolution ultra-high-quality synchrotron diffraction data on a few oligonucleotides whose crystals are exceptionally well-ordered and diffract to better than $1.0\text{ \AA}$, provides the opportunity for phasing methods that do not rely on any heavy atoms being required. Several pioneering studies have shown that “direct” methods, which employ mathematical relationships

between phases, may be used in favorable cases to compute phases (Han, 2001), and then electron-density maps from native structures.

Macromolecular crystal structures are normally optimized with respect to the diffraction data by nonlinear least-squares fitting procedures, which formally minimize the differences between observed and calculated models for the structure factors. This is the process of crystallographic refinement. When the diffraction data does not extend to atomic resolution, it is necessary to incorporate information from established stereochemical and structural features (such as bond lengths and angles, planar geometry of the DNA bases, preferred torsion angles). These are used to set up intramolecular constraints and restraints between them and so improve the initial models. One of the most widely used programs for macromolecular refinement, X-PLOR/CNS, uses empirical energy terms as part of the minimized function to ensure optimal intra- and intermolecular geometry. The technique of simulated annealing has been adopted from molecular dynamics as an effective way of refining structures when large-scale (>1 Å) atomic movements are required, since conventional least-squares methods are inherently incapable of effecting such large changes.

Oligonucleotide and oligonucleotide-protein crystals are heavily hydrated, with often over 50% solvent content. It is typical in medium-resolution structures for only a small fraction of these water molecules to be located in electron-density maps, largely because their high mobility smears their electron density to below the signal-to-noise level of these maps. The majority of water molecules reported in these structures are unsurprisingly the least mobile ones, which are directly hydrogen-bonded to the structure – these are the “first-shell” water molecules (see Chaps. 3 and 5 for detailed discussions of water arrangements in nucleic acid structures). The ways in which molecules pack in the crystal are sometimes of importance when examining structural features, since considerations of efficient packing can readily force parts of molecules to interact one with another by hydrogen bonding and van der Waals interactions, and consequently possibly modify some features of otherwise flexible conformation.

The quality and reliability of an oligonucleotide crystal structure are not straightforward to assess, especially for a noncrystallographer. Yet, judgments on these factors are critical when undertaking and using structural comparisons and analyses. The important crystallographic parameters of quality (R , R_{free}) have been outlined above. Of at least equal significance are the derived stereochemical features – examination of these is a reliable guide to quality (Das *et al.*, 2001).

Particular features to examine in a structure include:

- Close nonbonded intra- and intermolecular contacts that are less than the sum of the van der Waals radii of the atoms involved
- The distribution of values for torsion angles around single bonds. Eclipsed ($\sim 0^\circ$) values are indicative of problems in refinement
- Hydrogen bonds with distances appreciably outside the accepted ranges of ~ 2.7 – 3.2 Å
- Estimates of error in atomic positions
- The quality of the electron density for individual groups and atoms
- Values of atomic temperature factors, especially for water molecules

Checks on the purely structural features are equally applicable when examining structural models derived from NMR analyses. In practice all new crystal and NMR structures are rigorously checked for consistency when they are being deposited in the data bases, and any problems are drawn to the attention of the investigators. However it remains the case

that a significant number of older structures retain some problematic features that have not been corrected.

1.3 NMR Methods for Studying Nucleic Acid Structure and Dynamics

The underlying principle of nuclear magnetic resonance is the detection in a magnetic field of those atomic nuclei in a molecule, which have nuclear spin. Protons are abundant in nucleic acids and oligonucleotides, and fortunately have readily detectable spin signals. These signals, termed chemical shifts, are dependent on the shielding effect of neighboring protons, and thus can be used to determine the chemical environment of a proton once they can be unequivocally assigned as arising from particular atoms. Examples of highly characteristic, conformation-dependent chemical shifts are those arising from the protons on a deoxyribose sugar, which vary according to the pucker of the sugar (see Chap. 3). Other nuclei are less sensitive than protons, and have low natural abundance but the high field strength of magnets currently used in NMR studies are making ^{13}C and ^{15}N -enriched oligonucleotides amenable to detailed studies. NMR studies of oligonucleotides have been extensively used to examine interactions with proteins and small molecules, by monitoring characteristic changes in particular chemical shifts.

The magnetic interactions between a pair of protons gives rise to a NMR spin–spin coupling constant which is directly related to the dihedral angle between them, by the Karplus relationship. Hence measurement of coupling constants provides direct and reliable information on sugar puckers and on part of the backbone conformation (notably the glycosidic angle between sugar and base) in a nucleotide or oligonucleotide. Use of ^{13}C and ^{15}N labelled oligonucleotides enables coupling constants to be determined for otherwise inaccessible backbone torsion angles.

NMR methods enable structures to be determined in solution, largely by means of measurements of proton–proton coupling constants and through-space nuclear Overhauser effect (nOe) derived distances using 2D NMR methods. Solution-phase studies have the obvious advantage that molecules do not have to be crystallized, which is often the major (and highly frustrating!) limitation to the analysis of a macromolecule by X-ray crystallography. There is also the apparent advantage that a structure determined in solution is more relevant to physiological processes than an X-ray crystallographic study in the solid state. However, the two techniques should not be considered as alternatives. Rather they are complementary, providing distinct information. For example, NMR results emphasize the flexible nature of DNA and RNA molecules and the fact that individual groups such as sugars are dynamically in motion. It is notable that parallel observations of sequence-dependent effects in a number of oligonucleotide sequences have been reported from both crystallographic and NMR studies, although differences in detail are sometimes apparent.

There are a number of limitations to the accuracy and reliability of NMR methods as applied to nucleic acids. By contrast with crystallography, there is a limitation on the size of problem that can be analyzed in detail, due to the increase in the number of signals with molecular weight, and consequent overlap of chemical shifts. The nOe is only significant for protons, and so phosphate geometry is not directly defined by it. Hence nucleotide backbone conformations are in principle incompletely defined by standard ^1H NMR methods. Since a nOe intensity is proportional to the inverse sixth power of a proton–proton distance, it is a short-range effect, which is only significant

at distances less than 5–6 Å. Longer distances important in DNA structure, such as groove width, may not be derived directly. Most reliability from NMR experiments can be placed on particular features of DNA structure, especially base pairing, sugar pucker, and location of ligand binding sites. Detailed aspects of, in particular sequence-dependent structure, have been until recently a matter of considerable controversy. This is in large part because of the relatively small number of nOe data available compared to the several thousands of X-ray intensities from a typical medium resolution crystallographic analysis. The consequent under-determination of NMR-derived DNA structures means that their effective “resolution” is probably ~3–4 Å, with those parts of a structure providing the most nOe data being the most reliable. This situation is distinct from that for small (<25 kD) globular proteins, where the richness of nOe data arising from compact, closely packed amino-acid residues, provides for highly reliable and detailed NMR structure determination more equivalent to that from high-resolution crystallographic analyses.

A standard approach in NMR structure determination is to use restrained molecular dynamics methods and an assumed rough starting model. Commonly used programs are X-PLOR and CNS, thus sharing parameters and algorithms with crystallographic refinements. The distances established from individual nOe assignments are taken as constraints together with the NMR-derived conformational angles, to arrive at a plausible model or set of models. It is common practice to employ the nOe data in a semiquantitative manner, with the nOe signals being assigned to three groups. These correspond to long (e.g. 4–6 Å), medium (3–5 Å), and short (2–3.5 Å) interatomic distances. The accuracy of an NMR structure can be assessed by back-calculating the nOe intensities and comparing them with the observed, in a manner analogous to that used in crystallography. However the NMR “R factor” can be a less reliable guide since the number of observations is so much less than in crystallography, and flexibility in part of a structure may bias the R factor to give an apparently poor value. NMR studies produce an ensemble of related structures, rather than the single one from crystallography (Fig. 1.4). The distribution of particular elements of a structure is a consequence of the small number of experimental observations relative to the situation in a typical crystal structure. It also indicates the relative flexibility of different regions in a structure—typically in a nucleic acid the base pairs are the least flexible component compared to extrahelical loop regions.

1.4 Molecular Modelling and Simulation of Nucleic Acids

Crystallographic analyses provide a quasi-static view of molecular structure. The process of X-ray data acquisition from a single crystal, even at a high-flux synchrotron source, can take typically several minutes (although Laue methods can enable time-resolved diffraction data to be collected in the timescales of chemical and biochemical events). The vast majority of crystal structures provide a time-averaged picture of molecular motions about the low-energy structure in the crystal (which is typically >50% solvent). By contrast, molecular modeling techniques enable dynamic changes in structure and conformation to be calculated and visualized in terms of their effects on molecular energetics. These theoretical methods thus provide information complementary to the experimental techniques.

It is not feasible at present to compute conformational or energetic properties for significant lengths of nucleic acid sequence by *ab initio* quantum mechanics. Instead,

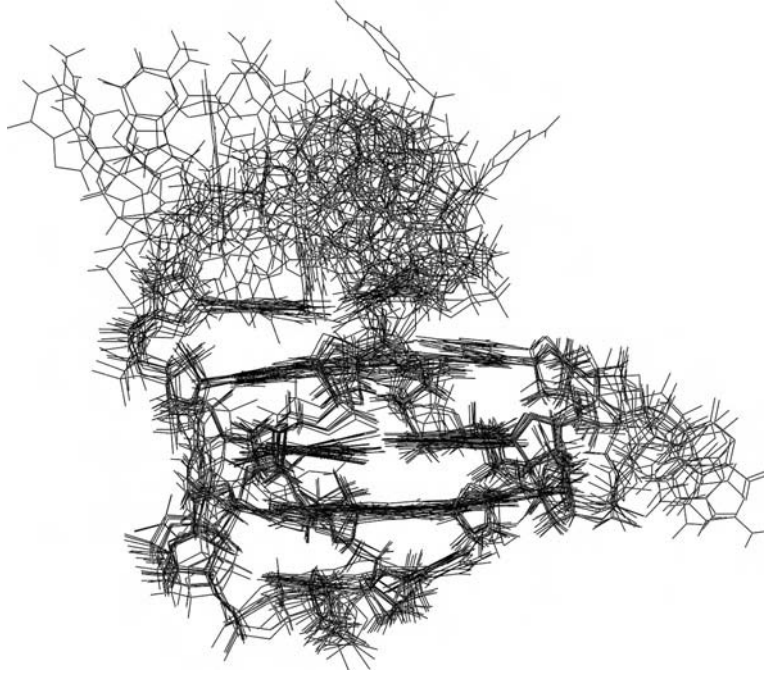


Figure 1.4 An ensemble of 10 structures from an NMR determination, of a DNA quadruplex formed by a sequence found in the human *bcl2* promoter (PDB entry no. 2F8U) (Dai *et al.*, 2006). Note the close correspondence of the bases in the various structures, contrasting with the diversity of conformation and flexibility shown by the extrahelical groups.

empirical force-field methods are widely used. These have been derived from experimental data that describe the energetics of a DNA or RNA molecule in terms of the sum of a number of factors:

$$\begin{aligned}
 V(r^N) = & \sum_{\text{bonds}} \frac{1}{2} k_b (l - l_0)^2 + \sum_{\text{angles}} \frac{1}{2} k_a (\theta - \theta_0)^2 \\
 & + \sum_{\text{torsions}} \frac{1}{2} V_n [1 + \cos(n\omega - \gamma)] \\
 & + \sum_{j=1}^{N-1} \sum_{i=j+1}^N \left\{ 4\epsilon_{i,j} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right\}
 \end{aligned}$$

- van der Waals nonbonded interactions, the 6–12 potential in the above formalism
- Bond length and angle distortions
- Barriers to rotation about single bonds

- Coulombic electrostatic contributions from full and partial electrostatic-potential-derived atomic charges
- Hydrogen-bonding (often incorporated as an implicit part of the electrostatic component)

The more recently developed nucleic acid force fields can also include the contributions of polarization effects, which can be especially important when examining the interactions of nucleic acids with drug or protein molecules.

The major empirical nucleic acid force fields have been incorporated into algorithms that can be used to minimize the conformation of a molecule with respect to its internal energy. This is the method of molecular mechanics, which in effect optimizes local low-energy minima. Much more extensive explorations of conformations can be made by molecular dynamics (MD), which applies Newton's equations of motion to an empirical force field, for all atoms in a molecule. This technique is computer-intensive, although even with current desktop workstations and high-end Linux-based PCs it is possible to undertake realistic simulations of molecular motions, with the inclusion of large numbers of calculated solvent molecules. The widespread availability of supercomputers and multiprocessor clusters has enabled an increasing number of simulations to be performed over thousands of picoseconds of molecular movements, with a number of studies of >10 nanoseconds (1 nanosecond = 1000 picosecond) duration. A very few have now reached the 1 microsecond limit, which is still faster than most biochemical events, but the rate of increase in computer power makes true biological simulations an attainable goal in the future.

Molecular dynamics can enable barriers between local energy minima to be traversed, unlike molecular mechanics. It is widely used in conjunction with distance geometry data (from nuclear Overhauser effects) to derive plausible DNA/RNA and nucleic acid-protein structures from these NMR measurements. The technique of simulated annealing is often employed, in which a structure is first simulated by molecular dynamics at a high temperature, when it is able to overcome high-energy barriers between conformations. The system is then gradually "cooled," when the most likely energy states become populated.

Solvent and counter-ions are normally routinely incorporated into dynamics simulations, their positions having been generated by Monte Carlo algorithms. The earlier use of a distance-dependent dielectric model to compensate for the lack of solvent often led to unstable structures during MD simulation. Even with the inclusion of a simple solvent model, care is needed to check that nucleic acid simulations remain stable due to the inability of the force field to adequately account for long-range electrostatic forces and solvation effects. This is typically seen in the breaking apart of structures after a period of simulation, which can be several hundred picoseconds. The use of Ewald and particle-mesh Ewald summation methods has led to a marked improvement in the stability of nucleic acid systems even in long timescale simulations. Validation against both NMR and X-ray crystallographic structures is an important activity, especially with the advent of a new generation of accurate and reliable experimental structures. Simulations with explicit solvent are computationally expensive, but use of the reliable generalized Born solvation model usually enables DNA to be modeled without the need for explicit water and counter-ions molecules to be present. This treatment can adequately account for solvent continuum effects and solute-solvent electrostatic polarization.

The use of molecular mechanics and dynamics methods have greatly increased in recent years, due not only to the ready availability of high-performance computing facilities, but also to the widespread availability of well-validated academic and commercially derived computer programs with graphical front-ends for the display of results. A number of modeling and simulation programs are in common use which have their force fields specifically parameterized for nucleic acids and their components. A detailed comparison has been made of simulations with different force fields applied to a B-DNA decamer system (Reddy, Leclerc, and Karplus, 2003). Disconcertingly, not all of the simulations resulted in the same overall B-type helix, although many parameters in these force fields have been updated since this study was undertaken (see, e.g., Soares *et al.*, 2005).

The most widely used and tested simulation programs are:

- AMBER (Assisted Model building and Energy Refinement), originally from the laboratory of the late P. A. Kollman, University of California, San Francisco. The force field incorporated in recent releases of this package, have been especially well-validated for nucleic acids (Case *et al.*, 2005; Cornell *et al.*, 1995; Wang *et al.*, 2004), and is generally taken to be the gold standard. The current version of the package is AMBER9. An updated AMBER force field is now available, that has been validated for improving the representation of some backbone geometry (Pérez *et al.*, 2007). AMBER is available from <http://www.amber.ucsf.edu>
- CHARMM (Chemistry at Harvard Macromolecular Mechanics), originally from the laboratory of M. Karplus, Harvard University. The forerunner of X-PLOR and CNS. The latter is available from <http://cns.csb.yale.edu/v1.1/>
- GROMOS (Groningen Molecular Simulation), (Soares *et al.*, 2005) developed in the laboratory of W. F. van Gunsteren and H. J. C. Berendsen, ETH Zurich. Available from <http://www.igc.ethz.ch/gromos/>
- JUMNA (Junction Minimization of Nucleic Acids) from the laboratory of Richard Lavery, Institut de Biologie Physico-Chimique, France. This program enables the torsion angles and helicoidal parameters in a structure to be varied, rather than standard x,y,z cartesian coordinates, and enables large regions of conformational space to be rapidly explored. Available from <http://www.ibpc.fr/UPR9080/>

1.5 Chemical, Enzymatic, and Biophysical Probes of Structure and Dynamics

Enzymes such as DNase I cleave the phosphodiester bond in a DNA duplex at every nucleotide position, although the cutting efficiency is markedly dependent on sequence, and by implication, on sequence-related structural features. Cleavage may be blocked by protein or drug binding. Hence DNase I can be used to determine sites of binding along a DNA sequence as well as to assess possible effects of particular sequences on DNA structure. Chemical cleaving agents such as hydroxyl radicals can give similar information. Since these are much smaller molecules than cleavage enzymes, their effects on DNA structure are less perturbing and sequence-dependent. Other types of chemical probe can attack specific base sites. These can be useful in defining the precise sites of protection resulting from drug or protein binding to a DNA sequence.

These footprinting methods have the important advantage over the fine-structure techniques of crystallography and NMR, of being applicable to long (up to several thousand base pair) DNA sequences, and thus of being more directly relevant to DNA in the cell. Hence, the use of chemical and enzymatic probes for DNA provides a way of obtaining at least some molecular-level data on otherwise inaccessible structural problems in DNA-protein and drug recognition. Footprinting can also be quantitated to provide kinetic and thermodynamic information at specific sites, although other techniques, notably surface plasmon resonance (SPR) and isothermal titration calorimetry (ITC) generally provide more accurate data.

Other structural methods that are beyond the scope of this book include atomic force microscopy, which is a powerful method of imaging large nucleic acids and protein–nucleic acid complexes. Single-molecule force measurement methods monitor events such as structural transitions, folding, and nucleic acid enzymology, providing information that complements and extends that from the fine-structure approaches described in this book.

1.6 Sources of Structural Data

The results of a crystal structure, fiber diffraction analysis or NMR study are most useful as a set of atomic coordinates. Those from fiber diffraction are available either in the primary literature or in various review chapters and compilations (see Chap. 3). Crystallographic coordinates may be obtained from a database, provided that they have been deposited in the first instance. This is no longer a problem since almost all journals now insist on deposition by publication, although authors still retain the right to up to a year's delay before public release of a data set. All available oligonucleotide crystal structures are in the successor to the Brookhaven Protein Data Bank, the Research Collaboration for Structural Biology (RCSB) Protein Data Bank (widely known as the PDB) (<http://www.rcsb.org>), and in the Macromolecular Structure Database at the European Bioinformatics Institute (EBI) (<http://www.ebi.ac.uk/msd/services.html>). Many depositions are accompanied by structure factor data. The PDB also contains a number of modeled structures. The Cambridge Crystallographic Database (which is primarily for small molecules), also contains coordinate data on a number of oligonucleotide structures (<http://www.ccdc.cam.ac.uk>). The Nucleic Acid Database (NDB) is a comprehensive relational database for nucleic acid crystallographic data, at Rutgers University, USA (Berman *et al.*, 1992) (<http://ndbserver.rutgers.edu>). It also provides a set of powerful tools for the comparative study of nucleic acid structural features, enabling detailed analysis of trends and features to be readily undertaken. Both the PDB and the NDB contain all deposited protein–nucleic acid structures, as well as many NMR-derived ones. Many NMR structures are also available in the BioMagRes-Bank Repository at <http://www.bmrb.wisc.edu>.

These databases are accessible via the internet from their primary sites, or from a number of mirror sites worldwide.

1.7 Visualization of Nucleic Acid Molecular Structures

Molecular structures are best viewed interactively on a computer screen, rather than as flat representations on a page. This book provides tables of Protein Data Bank and

Nucleic Acid Database identification codes, so that the reader can easily access and display individual structures, either by direct access to the PDB, NDB, or the EBI and the extensive visualization tools available in them, or by downloading coordinates from the PDB/NDB/EBI and inputting them into a molecular visualization or modeling program, such as one of those listed here.

A large number of such programs are now available, either commercially or as freeware. Many have been implemented on multiple platforms. Listed here are some recommended freeware programs that can be used on a PC (Windows or a Linux dialect) or a Macintosh desktop computer or laptop, and which are nucleic acid structure-friendly in that they can cope with nucleic acid residues and can also display nucleic acids in cartoon form. Several of these programs can directly read files from the PDB on the web.

1. **RasMol**, originally developed by Roger Sayle, was the first freely available PC molecular graphics program, and is now available from a number of sites, e.g. <http://www.openrasmol.org/>, and in a number of modifications, e.g. as Protein Explorer from <http://www.umass.edu/microbio/rasmol/>
2. **UCSF Chimera**, available from <http://www.cgl.ucsf.edu/chimera/>. Developed by the Computer Graphics Laboratory, University of California, San Francisco, USA (Couch *et al.*, 2006).
3. **PyMOL**, available from <http://pymol.sourceforge.net/>. Developed by DeLano Scientific, California, USA, available from <http://www.ebi.ac.uk/msd/index.html>
4. **VMD** (Visual Molecular Dynamics), available from <http://www.ks.uiuc.edu/Research/vmd/>. Developed by the Theoretical and Molecular Biophysics Group, University of Illinois at Urbana, Champaign, USA. (Humphrey, Dalke, and Schulten, 1996).
5. **FirstGlance in Jmol**, an open source viewer, which can be run either as a stand-alone program or as a web browser. Available from <http://molvis.sdsc.edu/fgij/index.htm> or <http://firstglance.jmol.org>

1.7.1 The Structures in This Book

Structures have been drawn using a range of different representations – sometimes more than one has been used in a single figure:

- Cartoons in which a phosphodiester backbone is represented by a ribbon, with the 5'→3' indicated by an arrow; bases and sugars shown as filled slabs
- Diagrams with bonds shown as lines or solid sticks
- Figures with bonds shown as solid sticks and atoms as small spheres
- van der Waals representations, having atoms drawn as spheres with radii set at their van der Waals values
- Surface representations, showing the solvent-excluded surface of a molecule

References

- Arnott, S. (1970). The geometry of nucleic acids. *Prog. Biophys. Mol. Biol.*, **21**, 265–319.
- Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S.-H., Srinivasan, A. R., and Schneider, B. (1992). The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.*, **63**, 751–759.

- Case, D. A., Cheatham, T. E. III, Darden, T., Gohlke, H., Luo, R., Merz, K. M. Jr, Onufriev, A., Simmerling, C., Wang, B., and Woods, R. J. (2005). The AMBER biomolecular simulation programs. *J. Comp. Chem.*, **26**, 1668–1688.
- Cornell, W. D., Cieplak, C. I., Bayly, I. R., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. (1995). A second generation force field for the simulation of proteins, nucleic acids and organic molecules. *J. Amer. Chem. Soc.*, **117**, 5179–5197.
- Couch, G. S., Hendrix, D. K., and Ferrin, T. E. (2006). Nucleic acid visualization with UCSF Chimera. *Nucleic Acids Res.*, **34**, e29.
- Dai, J., Chen, D., Jones, R. A., Hurley, L. H. and Yang, D. (2006). NMR solution structure of the major G-quadruplex structure formed in the human BCL2 promoter region. *Nucleic Acids Res.*, **34**, 5133–5144.
- Das, U., Chen, S., Fuxreiter, M., Vaguine, A. A., Richelle, J., Berman, H. M., and Wodak, S. J. (2001). Checking nucleic acid crystal structures. *Acta Crystallogr.*, **D57**, 813–828.
- Dauter, Z. and Adams, D. A. (2001). Anomalous signal of phosphorus used for phasing DNA oligomer: importance of data redundancy. *Acta Crystallogr.*, **D57**, 990–995.
- Ducruix, A. and Giegé, R. (Eds.). (1999). *Crystallisation of Nucleic Acids and Proteins*, 2nd Ed., Oxford University Press, Oxford.
- Han, G. W. (2001). Direct-methods determination of an RNA/DNA hybrid decamer at 1.15 Å resolution. *Acta Crystallogr.*, **D57**, 213–218.
- Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD – Visual Molecular Dynamics. *J. Molec. Graphics*, **14**, 33–38.
- Jiang, J., Sheng, J., Carrasco, N., and Huang, Z. (2007). Selenium derivatization of nucleic acids for crystallography. *Nucleic Acids Res.*, **35**, 477–485.
- Ke, A. and Doudna, J. A. (2004). Crystallization of RNA and RNA-protein complexes. *Methods*, **34**, 408–414.
- Pallan, P. S. and Egli, M. (2007a). Selenium modification of nucleic acids: preparation of oligonucleotides with incorporated 2'-SeMe-uridine for crystallographic phasing of nucleic acid structures. *Nature Protocols*, **2**, 647–651.
- Pallan, P. S. and Egli, M. (2007b). Selenium modification of nucleic acids: preparation of phosphoroselenoate derivatives for crystallographic phasing of nucleic acid structures. *Nature Protocols*, **2**, 640–646.
- Pérez, A., Marchán, I., Svozil, D., Sponer, J., Cheatham, T. E. III., Lughton, C. A., and Orozco, M. (2007). Refinement of the AMBER force field for nucleic acids: improving the description of α/γ conformers. *Biophys. J.*, **92**, 3817–3829.
- Reddy, S. Y., Leclerc, F., and Karplus, M. (2003). DNA polymorphism: a comparison of force fields for nucleic acids. *Biophys. J.*, **84**, 1421–1449.
- Salon, J., Sheng, J., Jiang, J., Chen, G., Caton-Williams, J., and Huang, Z. (2007). Oxygen replacement with selenium at the thymidine 4-position for the Se base pairing and crystal structure studies. *J. Amer. Chem. Soc.*, **129**, 4862–4863.
- Soares, T. A., Hunenberger, P. H., Kastenholtz, M. A., Krautler, V., Lenz, T., Lins, R. D., Oostenbrink, C., and van Gunsteren, W. F. (2005). An improved parameter set for the GROMOS force field. *J. Comp. Chem.*, **26**, 725–737.
- Tsui, V. and Case, D. A. (2000). Molecular dynamics simulations of nucleic acids with a generalized Born solvation model. *J. Amer. Chem. Soc.*, **122**, 2489–2498.
- Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004). Development and testing of a general AMBER force field. *J. Comp. Chem.*, **25**, 1157–1174.
- Wilds, C. J., Pattanayek, R., Pan, C., Wawrzak, Z., and Egli, M. (2002). Selenium-assisted nucleic acid crystallography: use of phosphoroselenoates for MAD phasing of a DNA structure. *J. Amer. Chem. Soc.*, **124**, 14910–14915.

Further Reading

General

- Blackburn, G. M., Gait, M. J., Loakes, D., and Williams, D. M. (Eds.) (2006). *Nucleic Acids in Chemistry and Biology*, 3rd Edition, The Royal Society of Chemistry, Cambridge, UK. An excellent survey of nucleic acid function, with some emphasis on structural aspects.
- Bloomfield, V. A., Crothers, D. M., and Tinoco, I., Jr. (2000). *Nucleic Acids. Structures, Properties and Functions*. University Science Books, Sausalito, California. An indispensable book for much background material, emphasizing the biophysical chemistry of nucleic acids, but covering many structural aspects as well.
- Calladine, C. R., Drew, H. R., Luisi, B., and Travers, A. A. (2004). *Understanding DNA*, 3rd Edition, Academic Press, London. An excellent account of selected areas of DNA structure, written with real understanding of the complexity of DNA structure and its biological consequences.

- Egli, M. (2004). Nucleic acid crystallography: current progress. *Curr. Opin. Struct. Biol.*, **8**, 580–591. A good recent summary of many of the topics covered in the first part of this book.
- Lilley, D. M. J. and Dahlberg, J. E. (Eds.) (1992). *Methods in Enzymology: DNA Structures*. Volumes 211, 212. Academic Press, San Diego. A comprehensive compilation of reviews on many aspects of DNA structure, emphasizing the methodologies of the early 1990s.
- Neidle, S. (Ed.) (1999). *Oxford Handbook of Nucleic Acid Structure*, Oxford University Press, Oxford. A detailed reference book on atomic-level DNA and RNA structures in the crystalline and solution states as studied by X-ray and NMR methods.
- Saenger, W. (1984). *Principles of Nucleic Acid Structure*. Springer-Verlag, Berlin. Dated in many areas, but still an important reference book, providing a wealth of background information and detail, and historic perspective.
- Sinden, R. R. (1994). *DNA Structure and Function*. Academic Press, San Diego.

The History of DNA Structure

The story of the discovery of the double-helical structure of DNA is in itself a minor industry, not least of the continuing controversies surrounding the discovery and its protagonists. This short selection of articles and books, taken together, provides most of the story. Brief reviews can also be found in the 50th anniversary Double Helix supplement published by *Nature* in 2003. The first two articles in this list have been written by experts in nucleic acid fiber diffraction who were involved in the important series of studies that rigorously defined the details of the fiber-diffraction structures in the decade after 1953.

- Arnott, S. (2006). Historical article: DNA polymorphism and the early history of the double helix. *Trends Biochem Sci.*, **31**, 349–354.
- Fuller, W. (2003). Who said helix? *Nature*, **424**, 876–878. A short but illuminating account of the discovery of the double-helix structure.
- Klug, A. (2004). The discovery of the double helix. *J. Mol. Biol.*, **335**, 3–26. A more detailed article, which includes material not previously generally available.
- Maddox, B. (2002). *Rosalind Franklin*. Harper Collins, London. An excellent biography that combines scientific accuracy with real insight.
- Olby, R. (1974). *The Path to the Double Helix*. Macmillan Press, London. An authoritative and detailed account of the history of the discovery of the double-helical structure and the background of earlier structural studies.
- Watson, J. D. (1968). *The Double Helix*. Weidenfeld and Nicolson, London. A highly personalized and entertaining account of the discovery of DNA.

More Detailed Accounts of Particular Topics can be Found in the Following

X-ray Crystallography

- Blow, D. M. (2002). *Outline of Crystallography for Biologists*, Oxford University Press, Oxford.
- Dauter, Z. (2005). Efficient use of synchrotron radiation for macromolecular diffraction data collection. *Prog. Biophys. Mol. Biol.*, **89**, 153–172.
- McPherson, A. (2003). *Introduction to Macromolecular Crystallography*, Wiley-Liss, New Jersey.
- McRee, D. E. (2000). *Practical Protein Crystallography*, 2nd Edition, Academic Press, San Diego.
- Rhodes, G. (2006). *Crystallography Made Crystal Clear*, 3rd Edition, Academic Press, San Diego. The website for this book is: <http://www.usm.maine.edu/~rhodes/CMCC/index.html>.
- Rossmann, M. G. and Arnold, E., (Eds.), (2001). *International Tables for X-ray Crystallography, Volume F, Crystallography of Biological Macromolecules*, Kluwer, Dordrecht.

NMR Methods

- Lane, A. N. (1995). Determination of fast dynamics of nucleic acids by NMR. *Methods in Enzymology*, **261**, 413–435.
- Patel, D. J., Shapiro, L., and Hare, D. R. (1987). DNA and RNA: NMR studies of conformations and dynamics in solution. *Quart. Rev. Biophys.*, **20**, 35–112.
- James, T. L. (Ed.), (2005). Nuclear magnetic resonance of biological macromolecules. *Methods in Enzymology*, Volume 394.
- Latham, M. P., Brown, D. J., McCallum, and Pardi, A. (2005). NMR methods for studying the structure and dynamics of RNA. *Chembiotech.*, **6**, 1492–1505.
- Wemmer, D. E. (1991). The applicability of NMR methods to solution structure of nucleic acids. *Curr. Opin. Struct. Biol.*, **1**, 452–458.

- Wüthrich, K. (1986). *NMR of Proteins and Nucleic Acids*, John Wiley, New York. *Molecular modelling and simulation methods*.
- Beveridge, D. L. and McConnell, K. J. (2000). Nucleic acids: theory and computer simulation. *Curr. Opin. Struct. Biol.*, **10**, 182–196.
- Cheatham, T. E. III. (2004). Simulation and modeling of nucleic acid structure, dynamics and interactions. *Curr. Opin. Struct. Biol.*, **14**, 360–367.
- Lavery, R., Zakrzewska, K., and Sklenar, H. (1995). JUMNA (junction minimization of nucleic acids). *Comput. Phys. Commun.*, **91**, 135–158.
- MacKerell, A. D., Jr. (2004). Empirical force fields for biological macromolecules: overview and issues. *J. Comput. Chem.*, **25**, 158–1604.
- McCammon, J. A. and Harvey, S. C. (1987). *Dynamics of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge, UK.
- van Gunsteren, W. F. and Berendsen, H. J. C. (1990). Computer simulation of molecular dynamics: methodology, applications, and perspectives in chemistry. *Ang. Chem. Intern. Ed.*, **29**, 992–1023.

Chemical, Enzymatic, Biophysical, and Single-Molecule Probes of Nucleic Acid Structure

- Bockelman, U. (2004). Single-molecule probes of nucleic acid structure. *Curr. Opin. Struct. Biol.*, **14**, 368–373.
- Fox, K. R., (Ed.), (1997). *Drug-DNA Interaction Protocols*. Humana Press, New Jersey.
- Tullius, T. D. and Greenbaum, J. A. (2005). Mapping nucleic acid structure by hydroxyl radical cleavage. *Curr. Opin. Chem. Biol.*, **9**, 127–134.
- Wilson, W. D. (2002). Analyzing biomolecular interactions. *Science*, **295**, 2103–2104.

The Nucleic Acid and Protein Databases

- Berman, H. M., Gelbin, A., and Westbrook, J. (1996). Nucleic acid crystallography: a view from the nucleic acid database. *Prog. Biophys. Molec. Biol.*, **66**, 255–288.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N. Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

The Building-Blocks of DNA and RNA

2.1 Introduction

Chemical degradation studies in the early years of the twenty-first century on material extracted from cell nuclei established that the high molecular-weight “nucleic acid” was actually composed of individual acid units, termed nucleotides. Four distinct types were isolated – guanylic, adenylic, cytidylic and thymidylic acids. These could be further cleaved to phosphate groups and four distinct nucleosides. The latter were subsequently identified as consisting of a deoxypentose sugar and one of four nitrogen-containing heterocyclic bases. Thus, each repeating unit in a nucleic acid polymer comprises these three units linked together – a phosphate group, a sugar, and one of the four bases.

The bases are planar aromatic heterocyclic molecules and are divided into two groups – the pyrimidine bases thymine and cytosine and the purine bases adenine and guanine. Their major tautomeric forms are shown in Fig. 2.1. Thymine is replaced by uracil in ribonucleic acids, which also have an extra hydroxyl group at the 2' position of their (ribose) sugar groups. The standard nomenclature for the atoms in nucleic acids, as approved by the International Union of Biochemistry, is shown in Figs. 2.1. and 2.2. Accurate bond length and angle geometries for all bases, nucleosides and nucleotides have been well established by X-ray crystallographic analyses. Structural surveys (Clowney *et al.*, 1996; Gelbin *et al.*, 1996) have calculated mean values for these parameters (which define their equilibrium values) from the most reliable structures in the Cambridge and Nucleic Acid Databases. These have been incorporated in several implementations of the AMBER and CHARMM force fields widely used in molecular mechanics and dynamics modelling, and in a number of computer packages for both crystallographic and NMR structural analyses (Parkinson *et al.*, 1996). Accurate crystallographic analyses, at very high resolution, can also directly yield quantitative information on the electron-density distribution in a molecule, and hence on individual partial atomic charges. These charges for nucleosides have hitherto been obtained by ab initio quantum mechanical calculations, but are also available experimentally for all four DNA nucleosides (Pearlman and Kim, 1990).

Individual nucleoside units are joined together in a nucleic acid in a linear manner, through phosphate groups attached to the 3' and 5' positions of the sugars (Fig. 2.2). Hence the full repeating unit in a nucleic acid is a 3',5'-nucleotide.

Nucleic acid and oligonucleotide sequences use single-letter codes for the five unit nucleotides – A, T, G, C and U. The two classes of bases can be abbreviated as

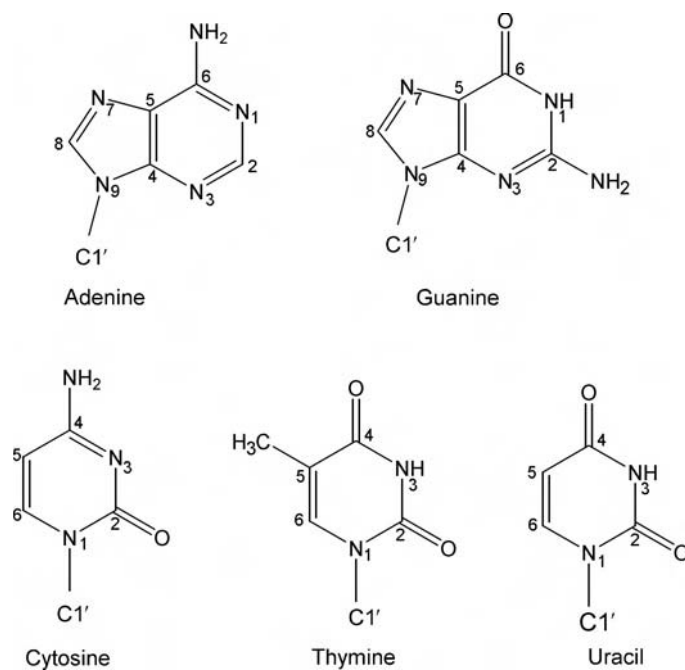


Figure 2.1 The five bases of DNA and RNA.

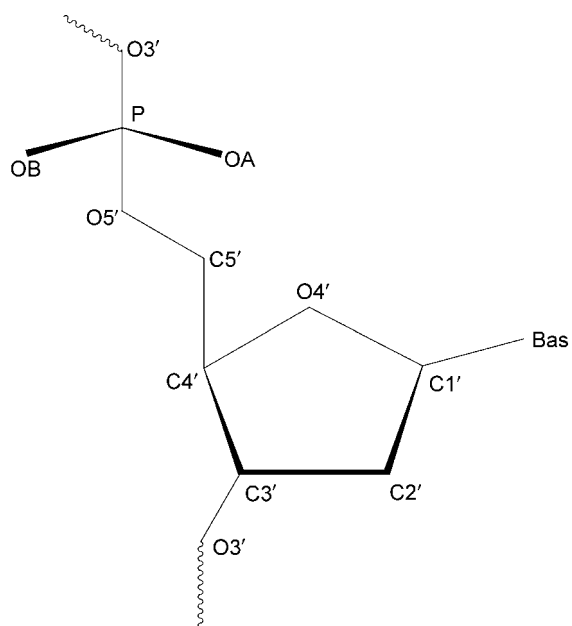


Figure 2.2 The organization of repeating units in a polynucleotide chain.

Y (pyrimidine) and R (purine). Phosphate groups are usually designated as p. A single oligonucleotide chain is conventionally numbered from the 5' end, for example, ApG-pCpTpTpG has the 5' terminal adenosine nucleoside, with a free hydroxyl at its 5' position and thus the 3' end guanosine has a free 3' terminal hydroxyl group. The letter p, to denote intervening phosphate groups, is sometimes omitted when a sequence is written down. Chain direction is sometimes emphasized with 5' and 3' labels. Thus an anti-parallel double-helical sequence can be written as:

5'CpGpCpGpApApTpTpCpGpCpG
 3'GpCpGpCpTpTpApApGpCpGpC
 or simply as
 CGCGAATTCGCG

Structural publications on DNA usually prefix a sequence with "d", as in d(CGAT), to emphasize that the oligonucleotide is a deoxyribose one rather than being an oligoribonucleotide. The prefix "r" to denote ribonucleosides, ribonucleotides and their oligomers, is often used.

The bond between sugar and base is known as the glycosidic bond. Its stereochemistry is important. In natural nucleic acids the glycosidic bond is always β , that is the base is above the plane of the sugar when viewed onto the plane and therefore on the same face of the plane as the 5' hydroxyl substituent (Fig. 2.3a). The absolute stereochemistry of other substituent groups on the deoxyribose sugar ring of DNA is defined such that when viewed end-on with the sugar ring oxygen atom O4' at the rear (Fig. 2.4a), the hydroxyl group at the 3' position is below the ring and the hydroxymethyl group at the 4' position is above it. A unit nucleotide can have its

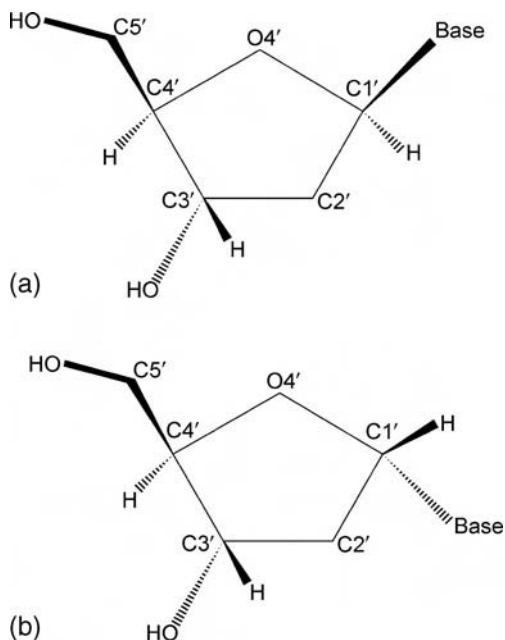


Figure 2.3 (a) The stereochemistry of a natural β -nucleoside. Solid bonds are coming out of the plane of the page, toward the reader. Dashed bonds are going away from the reader. (b) The stereochemistry of an α -nucleoside.

phosphate group attached either at the 3' or 5' ends, and is thus termed either a 3'- or a 5'- nucleotide. It is chemically possible to construct α -nucleosides and from them α -oligonucleosides, which have their bases in the "below" configuration relative to the sugar rings and their other substituents (Fig. 2.3b). These are much more resistant to nuclease attack than standard natural β -oligomers and have been used as antisense oligomers to mRNAs on account of their superior intracellular stability.

2.2 Base Pairing

The realization that the planar bases can associate in particular ways by means of hydrogen bonding was a crucial step in the elucidation of the structure of DNA. The important early experimental data of Chargaff showed that the molar ratios of adenine:thymine and cytosine:guanine in DNA were both unity. This led to the proposal by Watson and Crick that in each of these pairs the purine and pyrimidine bases are held together by specific hydrogen bonds, to form planar base pairs. In native, double-helical DNA the two bases in a base pair necessarily arise from two separate strands of DNA (with intermolecular hydrogen bonds) and so hold the DNA double helix together (Watson and Crick, 1953).

The adenine:thymine (A•T) base pair has two hydrogen bonds compared to the three in a guanine:cytosine (G•C) one (Fig. 2.4). (The initial Watson–Crick model assigned just two hydrogen bonds to the G•C pair, and the third hydrogen bond emerged some three years later (Wain-Hobson, 2006)). Fundamental to the Watson–Crick arrangement is that the sugar groups are both attached to the bases on the same side of the base pair. As will be seen in Chap. 3, this defines the mutual positions of the two sugar-phosphate strands in DNA itself. The two base pairs are required to be almost identical in dimensions by the Watson–Crick model. High resolution (0.8–0.9 Å) X-ray crystallographic analyses of the ribodinucleoside monophosphate duplexes r(GpC) and r(ApU) by A. Rich and colleagues in the early 1970s (Seeman *et al.*, 1976; Rosenberg *et al.*, 1976) established accurate geometries for these A•T and G•C base pairs (Table 2.1). These structure determinations showed that there are only small differences in size between the two types of base pairings, as indicated by the

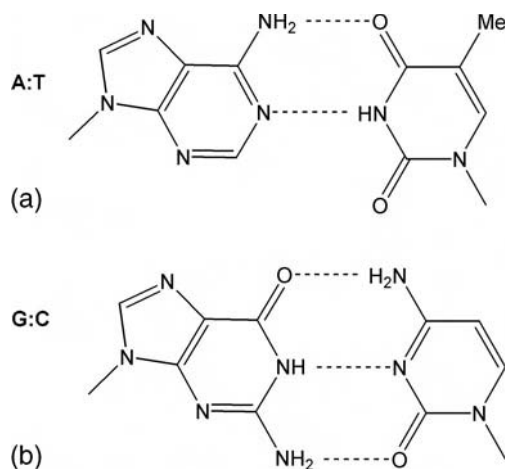


Figure 2.4 (a) A•T and (b) G•C base pairs, showing Watson–Crick hydrogen bonding.

Table 2.1 Hydrogen-Bond Distances (in Å) in Watson–Crick Base Pairs (Seeman *et al.*, 1976; Rosenberg *et al.*, 1976), in the Crystalline State. Estimated Standard Deviations are in Parentheses

U•A:	N3-H.....N1	2.835(8)
	O4.....H-N6	2.940(8)
C•G:	O2.....H-N2	2.86(1)
	N3.....H-N1	2.95(1)
	N4-H.....O6	2.91(1)

distance between glycosidic carbon atoms in a base pair. The C1'–C1' distance in the G•C base pair structure is 10.67 Å, and 10.48 Å in the A•U-containing dinucleoside.

Theoretical studies on base pairing have enabled estimates to be made of the extra stability conferred on a G•C base pair by the extra hydrogen bond. Molecular dynamics simulations of both base pairs in an aqueous environment have given the free energies of A•T and G•C base pairs as -4.3 and -5.8 kcal/mol (Stofer, Chipot, and Lavery, 1999). An alternative, reversed Watson–Crick A•T base pairs involving a 180° rotation of one base, had similar stability. A more recent estimate, using *ab initio* quantum mechanical methods (Sponer, Jurecka and Hobza, 2004), has shown broad agreement with values calculated from the AMBER force field. This reinforces the reliability of the force-field approaches that are widely used and which emphasize van der Waals and electrostatic point-centered charge contributions, even though they do not take account of polarization and charge transfer effects. The interaction enthalpies of hydrated A•T and G•C base pairs have been calculated as 14.0 kcal/mol and 27.0 kcal/mol respectively (i.e. 7–9 kcal/mol per hydrogen bond), using AMBER together with density functional theory calculations (Liu, Wyttenbach, and Bowers, 2006). Other A•T base pair arrangements have been predicted by theory (Gould and Kollman, 1994) to be more stable than Watson–Crick pairing. However these alternatives would require significant changes in backbone conformation in order to be accommodated within a duplex. It is striking that these alternatives have not been observed in normal duplex DNA, suggesting that the requirements for optimal base stacking and minimal backbone distortions greatly favor the standard Watson–Crick arrangement.

2.3 Base and Base Pair Flexibility

The individual bases in a nucleic acid are flat, but base pairs (and consecutive bases on an individual strand), can show considerable flexibility. This flexibility is to some extent dependent on the nature of the bases and base pairs themselves, but is more related to their base-stacking environments. Thus descriptions of base morphology have become important in describing and understanding many sequence-dependent features and deformations of nucleic acids. The former features are often considered primarily at the dinucleoside local level, whereas longer-range effects such as helix bending, can also be analyzed at a more global level.

A number of rotational and translational parameters have been devised to describe these geometric relations between bases and base pairs, which were originally defined (10) in 1989 (the “Cambridge Accord”). These definitions, together with the

Cambridge Accord sign conventions, are given later. Confusingly, two distinct types of approaches have been reported in the literature to calculate these parameters (Lu, Babcock, and Olson, 1999) – the Cambridge Accord did not define a single unambiguous convention for their calculation. In one approach, the parameters are defined with respect to a global helical axis, which need not be linear. The other uses a set of local axes, one per dinucleotide step. Also, a variety of definitions of local and global axes have been used. The overall effect for most undistorted structures is fortunately that only a minority of parameters appear to have very different values depending on the method of calculation, using a number of the widely available programs (see later). Before detailing these, we introduce the parameters themselves. The major and minor grooves are the indentations in nucleic acid double helices formed as a consequence of the asymmetry of base pairs. The C1'–N9 (purine) and C1'–N1 (pyrimidine) base-sugar bonds are by convention on the minor-groove side, so that the C6/N7 (purine) and C4 (pyrimidine) base atoms and their substituents are on the major-groove side. The grooves and their characteristics are described further in Chap. 3.

(i) For individual base pairs:

- (a) **Propeller twist** (ω) between bases is the dihedral angle between normals to the bases, when viewed along the long axis of the base pair (Fig. 2.5). The angle has a negative sign under normal circumstances, with a clockwise rotation of the nearer base when viewed down the long axis. The long axis for a purine–pyrimidine base pair is defined as the vector between the C8 atom of the purine and the C6 of a pyrimidine in a Watson–Crick base pair. Analogous definitions can be applied to other nonstandard base pairings in a duplex including purine–purine and pyrimidine–pyrimidine ones.
- (b) **Buckle** (κ) is the dihedral angle between bases, along their short axis, after propeller twist has been set to 0° (Fig. 2.6). The sign of buckle is defined as

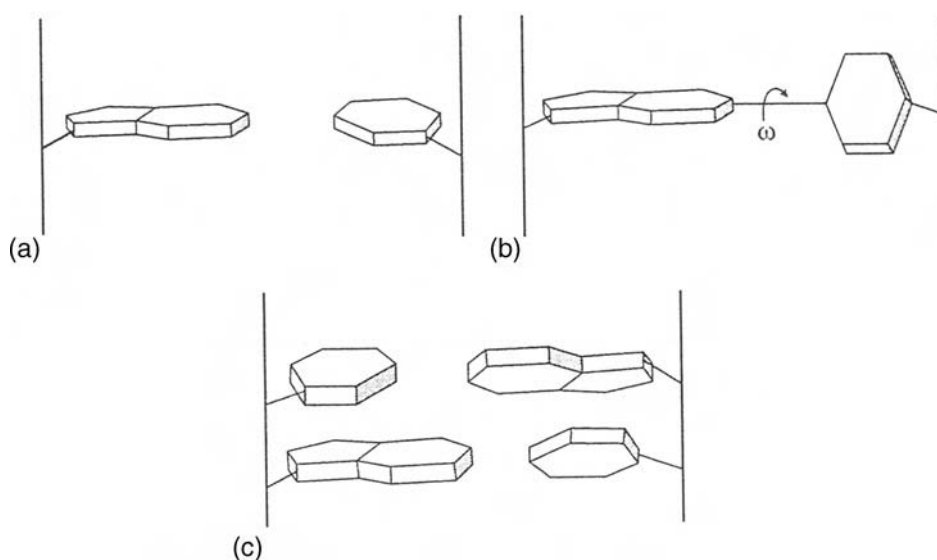


Figure 2.5 Schematic views of a base pair with (a) zero and (b) high propeller twist. Panel (c) shows the effects of propeller twist on two successive base pairs.

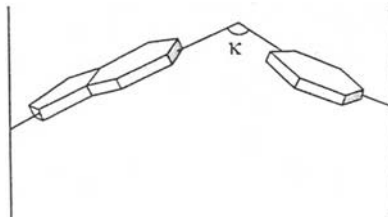


Figure 2.6 A schematic view of base-pair buckle.

positive if the distortion is convex in the direction $5' \rightarrow 3'$ of strand 1. The change in buckle for succeeding steps, termed **cup**, has been found to be a useful measure of changes along a sequence. Cup is defined as the difference between the buckle at a given step, and that of the preceding one.

- (c) **Inclination** (η) is the angle between the long axis of a base pair and a plane perpendicular to the helix axis. This angle is defined as positive for right-handed rotation about a vector from the helix axis toward the major groove.
- (d) **X and Y displacements** define translations of a base pair within its mean plane in terms of the distance of the midpoint of the base pair long axis from the helix axis. X displacement is toward the major-groove direction, when it has a positive value. Y displacement is orthogonal to this, and is positive if toward the first nucleic acid strand of the duplex.
- (ii) For base pair steps:
 - (e) **Helical twist** (Ω) is the angle between successive base pairs, measured as the change in orientation of the $C1'-C1'$ vectors on going from one base pair to the next, projected down the helix axis (Fig. 2.7). For an exactly repetitious double helix, helical twist is $360^\circ/n$, where n is the unit repeat defined above.
 - (f) **Roll** (ρ) is the dihedral angle for rotation of one base pair with respect to its neighbor, about the long axis of the base pair. A positive roll angle opens up a base pair step toward the minor groove (Fig. 2.8). **Tilt** (τ) is the corresponding dihedral angle along the short (i.e. x -axis) of the base pair.
 - (g) **Slide** is the relative displacement of one base pair compared to another, in the direction of nucleic acid strand one (i.e. the Y displacement), measured between the midpoints of each C6–C8 base pair long axis.

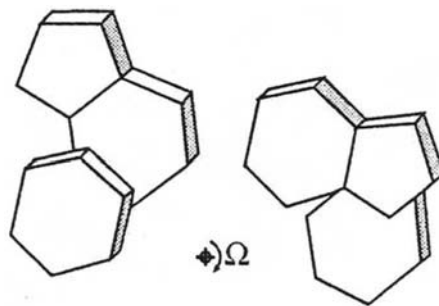


Figure 2.7 View down two successive base pairs, showing the helical twist angle between them.

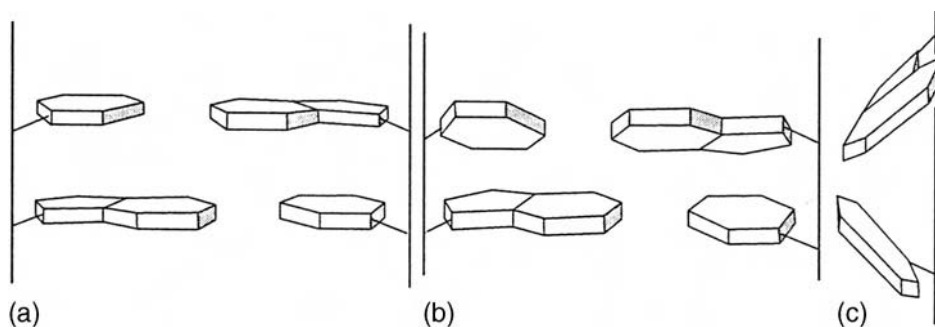


Figure 2.8 Views of two successive base pairs, (a) with 0° roll angle between them, and (b) with a positive roll angle. Panel (c) shows a view of positive roll along the long axis of the base pairs.

A standardized coordinate reference frame for the calculation of these parameters has been proposed (Lu and Olson, 1999), and have been endorsed by the successor to the Cambridge Accord, the 1999 Tsukuba Accord (Olson *et al.*, 2001). Details of the reference frame are also available on the Nucleic Acid Database web site at http://ndbserver.rutgers.edu/standards/supl_inf.html.

This reference frame is unambiguous and has the advantage of being able to produce values for the majority of local base-pair and base-step parameters which are almost independent of the algorithm used. A notable exception is rise, which is especially sensitive to the definition of origin and to small changes in buckle and roll. The right-handed reference frame used is shown in Fig. 2.9. It has the x -axis directed toward the major groove along the pseudo two-fold axis of an idealized Watson–Crick base pair (shown as $\bullet$). The y -axis is along the long axis of the base pair, parallel to the $C1'-C1'$ vector. The position of the origin is clearly dependent on the geometry of the bases and the base pair. These have been taken from the published compilations (Clowney *et al.*, 1996; Gelbin *et al.*, 1996).

The most widely-used computer programs for base and base-pair morphology calculations, all of which use the above parameter definitions, are:

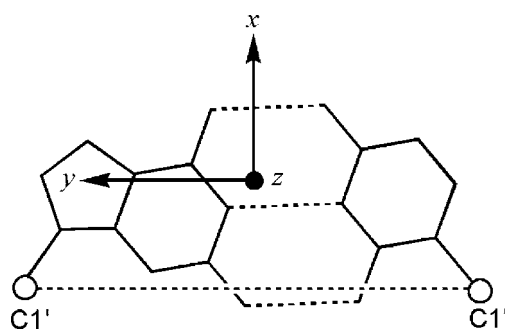


Figure 2.9 Reference frame for idealized helical DNA, showing the axis origin at ($\bullet$), along the pseudo two-fold axis between two successive base pairs.

- *CURVES* (Lavery and Sklenar, 1988, 1989), which calculates an optimized global helix axis which can be (and invariably is), curved. This is especially useful for irregular structures. The reference frame is defined at the base rather than the base pair level. Local parameters can also be obtained with this program, and irregular nonstandard base-pairing arrangements can be accommodated.
- *FREEHELIX*, the successor to the original helical parameter program NEWHELIX, calculates an optimum linear helical axis, so that the derived parameters are basically global ones (Dickerson, 1998). Local bending can also be calculated (Goodsell and Dickerson, 1994).
- *RNA* (Babcock, Pednault, and Olson, 1994) uses the reference frame subsequently adopted by the Tsukuba Accord, and calculates local helical parameters. It can also be used for nonstandard bases.
- *CEHS* (Lu, El Hassan, and Hunter, 1997) also focuses on local parameters, and like *FREEHELIX*, uses the C6–C8 base-pair vector as the y-axis for the reference frame. Roll and tilt are commuted into a single variable RollTilt (Γ).

Readers of crystallographic or NMR structure analysis literature should be aware of whether local or global axis definitions have been used in parameter calculation. Ambiguities can be clarified by use of the NDB, which can produce tables of base helical parameters calculated with all of these methods.

2.4 Sugar Puckers

The five-membered deoxyribose sugar ring in DNA is inherently nonplanar. This nonplanarity is termed puckering. The precise conformation of a deoxyribose ring can be completely specified by the five endocyclic torsion angles within it (Fig. 2.10). The ring puckering arises from the effect of nonbonded interactions between substituents at the four ring carbon atoms – the energetically most stable conformation for the ring has all substituents as far apart as possible. Thus different substituent atoms would be expected to produce differing types of puckering. The puckering can be described by either

- A simple qualitative description of the conformation in terms of atoms deviating from ring coplanarity; or
- Precise descriptions in terms of the ring internal torsion angles.

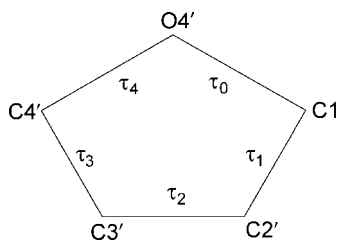


Figure 2.10 The five internal torsion angles in a ribose ring.

In principle, there is a continuum of interconvertible puckers, separated by energy barriers. These various puckers are produced by systematic changes in the ring torsion angles. The puckers can be succinctly defined by the parameters P and τ_m (Altona and Sundaralingam, 1972). The value of P , the phase angle of pseudorotation, indicates the type of pucker since P is defined in terms of the five torsion angles τ_0 – τ_4 :

$$\tan P = \frac{(\tau_4 + \tau_1) - (\tau_3 + \tau_0)}{2 * \tau_2 * (\sin 36^\circ + \sin 72^\circ)}$$

and the maximum degree of pucker, τ_m , by

$$\tau_m = \frac{\tau_2}{\cos P}$$

The pseudorotation phase angle can take any value between 0° and 360° . If τ_2 has a negative value, then 180° is added to the value of P . The pseudorotation phase angle is commonly represented by the pseudorotation wheel, which indicates the continuum of ring puckers (Fig. 2.11). Values of τ_m indicate the degree of puckering of the ring; typical experimental values from crystallographic studies on mononucleosides are in the range 25 – 45° . The five internal torsion angles are not independent of each other, and so to a good approximation any one angle τ_j can be represented in terms of just two variables:

$$\tau_j = \tau_m \cos[P + 0.8\pi(j - 2)]$$

An on-line interactive facility (PROSIT) is now available to calculate pseudorotation parameters (Sun *et al.*, 2004) at <http://www.cactus.nci.nih.gov/prosit/>.

A large number of distinct deoxyribose ring pucker geometries have been observed experimentally, by X-ray crystallography and NMR techniques. When one ring atom is out of the plane of the other four, the pucker type is an envelope one. More commonly, two atoms deviate from the plane of the other three, with these two either

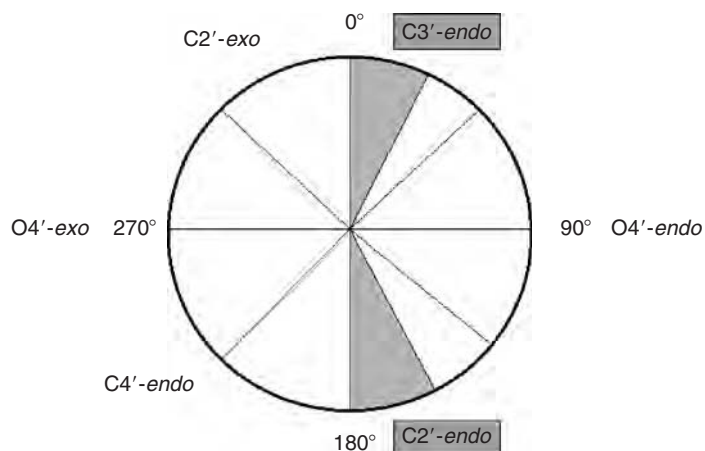


Figure 2.11 The pseudorotation wheel for a deoxyribose sugar. The shaded areas indicate the preferred ranges of the pseudorotation angle for the two principal sugar conformations.

side of the plane. It is usual for one of the two atoms to have a larger deviation from the plane than the other, resulting in a twist conformation. The direction of atomic displacement from the plane is important. If the major deviation is on the same side as the base and C4'–C5' bond, then the atom involved is termed *endo*. If it is on the opposite side, it is called *exo*. The most commonly observed puckers in crystal structures of isolated nucleosides and nucleotides are either close to C2'-*endo* or C3'-*endo* types (Fig. 2.12a,b). In practice, these pure envelope forms are rarely observed,

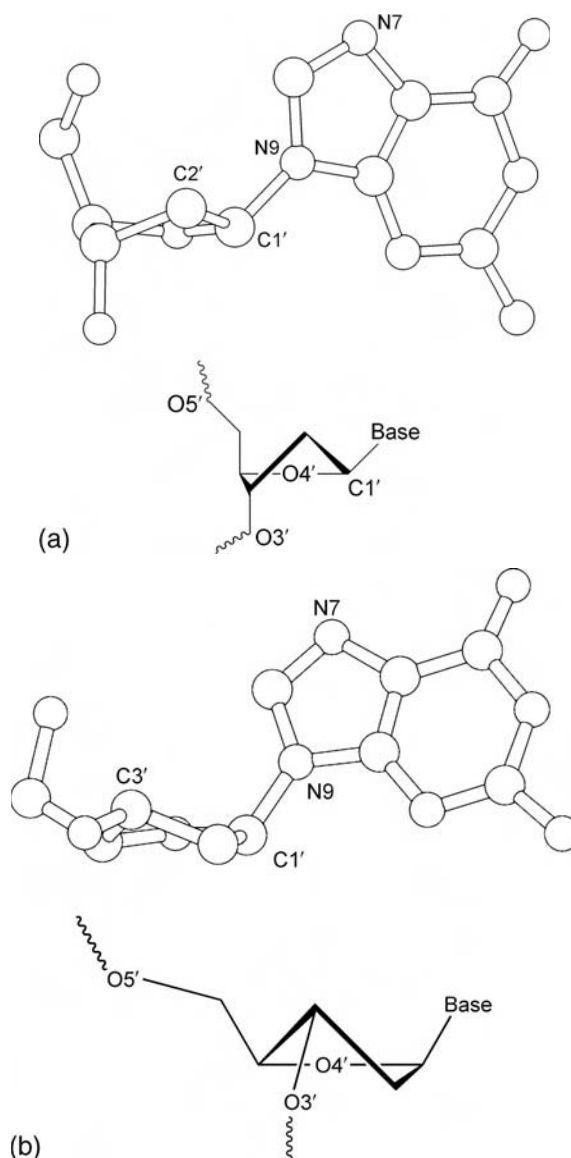


Figure 2.12 (a) C2'-*endo* sugar puckering for a guanosine nucleoside, viewed along the plane of the sugar ring. (b) C3'-*endo* sugar puckering for a guanosine nucleoside, with the sugar viewed in the same direction as in panel (a).

largely because of the differing substituents on the ring. Consequently the puckers are then best described in terms of twist conformations. When the major out-of-plane deviation is on the *endo* side, there is a minor deviation on the opposite, *exo* side. The convention used for describing a twist deoxyribose conformation is that the major out-of-plane deviation is followed by the minor one, for example C2'-*endo*, C3'-*exo*. The C2'-*endo* family of puckers have P values in the range of 140–185°; in view of their position on the pseudorotation wheel, they are sometimes termed S (south) conformations. The C3'-*endo* domain has P values in the range of –10 to +40°, and its conformation is termed N (north). These more geographical terms are mostly used by NMR spectroscopists.

The pseudorotation wheel implies that deoxyribose puckers are free to interconvert. In practice, there are energy barriers between major forms. The exact size of these barriers has been the subject of considerable study (Olson, 1982a; Olson and Sussman, 1982). The consensus is that the barrier height is dependent on the route around the pseudorotation wheel. For interconversion of C2'-*endo* to C3'-*endo* the preferred pathway is via the O4'-*endo* state, with a barrier of 2–5 kcal/mole found from an analysis of a large body of experimental data (Olson and Sussman, 1982), and a somewhat smaller (potential energy) value of 1.5 kcal/mole from a molecular dynamics study (Harvey and Prabhakaran, 1986). The former value, being an experimental one, represents the total free energy for interconversion. These values also broadly concur with those found by application of various *ab initio* methods, which have given values of 2.5–4.0 kcal/mole (Foloppe, Nilsson, and MacKerell, 2001).

Relative populations of puckers can be monitored directly by NMR measurements of the ratio of coupling constants between H1'–H2' and H3'–H4' protons. These show that in contrast to the “frozen-out” puckers found in the solid state structures of nucleosides and nucleotides, there is rapid interconversion in solution. Nonetheless, the relative populations of the major puckers are dependent on the type of base attached. Purines show a preference for the C2'-*endo* pucker conformational type whereas pyrimidines favor C3'-*endo*. Deoxyribose nucleosides are primarily (>60%) in the C2'-*endo* form and ribonucleosides favor C3'-*endo*. The latter are significantly more restricted in their mobility; this has significance for the structures of oligoribonucleotides (see Chap. 6). These differences in puckering equilibria and hence in their relative populations in solution and in molecular dynamics simulations, are reflected in the patterns of puckers found in surveys of crystal structures (Murray-Rust and Motherwell, 1978). Again, this is a demonstration of the complementarity of information provided by the different structural techniques. Sugar pucker preferences have their origin in the nonbonded interactions between substituents on the sugar ring, and to some extent on their electronic characteristics. For example, the C3'-*endo* pucker (Fig. 2.7b) would have hydroxyl substituents at the 2' and 3' positions further apart than with C2'-*endo* pucker; hence the preference of the former by ribonucleosides.

Numerous crystallographic and NMR studies have found correlations between sugar pucker and several backbone conformational variables, both in isolated nucleosides/nucleotides and in oligonucleotide structures. These are discussed later in this chapter. Changes in sugar pucker are important determinants of oligo- and polynucleotide structure because they can alter the orientation of C1', C3' and C4' substituents, resulting in major changes in backbone conformation and overall structure, as indeed is found (Chap. 3).

2.5 Conformations About the Glycosidic Bond

The glycosidic bond links a deoxyribose sugar and a base, being the C1'–N9 bond for purines and the C1'–N1 bond for pyrimidines. The torsion angle χ around this single bond can in principle adopt a wide range of values, although as will be seen, structural constraints result in marked preferences being observed. Glycosidic torsion angles are defined in terms of the four atoms:

O4'–C1'–N9–C4 for purines

O4'–C1'–N1–C2 for pyrimidines

Theory has predicted two principal low-energy domains for the glycosidic angle, in accord with experimental findings for a large number of nucleosides and nucleotides. The *anti* conformation has the N1, C2 face of purines and the C2, N3 face of pyrimidines directed away from the sugar ring (Fig. 2.13a) so that the hydrogen atoms attached to C8 of purines and C6 of pyrimidines, are lying over the sugar ring. Thus, the Watson–Crick hydrogen-bonding groups of the bases are directed away from the sugar ring. These orientations are reversed for the *syn* conformation, with these hydrogen-bonding groups now oriented toward the sugar and especially its O5' atom (Fig. 2.13b). A number of crystal structures of *syn* purine nucleosides have found hydrogen bonding between the O5' atom

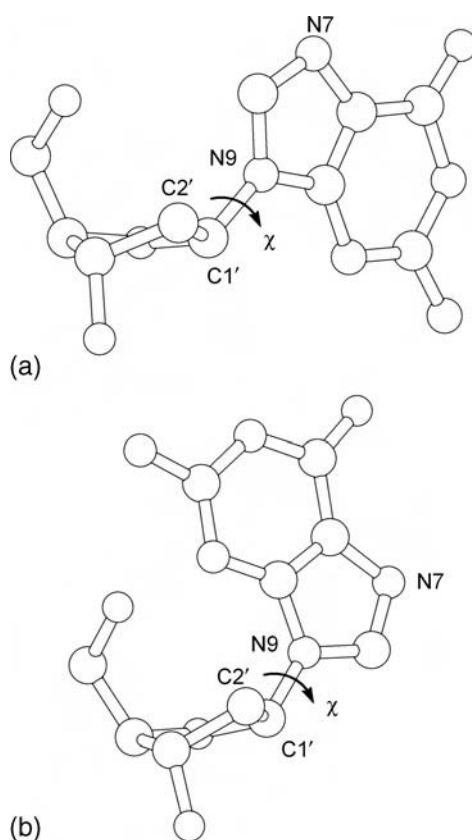


Figure 2.13 (a) A guanosine nucleoside with the glycosidic angle χ set in an *anti* conformation. (b) A guanosine nucleoside in the *syn* conformation.

Table 2.2 Energy Differences (in kcal/mole) Between Glycosidic Angle Conformers for B-form DNA Nucleotides, Calculated with the AMBER Program

Nucleotides	Energy difference
A(<i>anti</i>) > A(<i>syn</i>)	0.3
T(<i>anti</i>) > T(<i>syn</i>)	1.7
G(<i>syn</i>) > G(<i>anti</i>)	3.3
C(<i>anti</i>) > C(<i>syn</i>)	1.8

and the N3 base atom, which would stabilize this conformation. Otherwise, for purines, the *syn* conformation is slightly less preferred than the *anti*, on the basis of fewer nonbonded steric clashes in the latter case. The principal exceptions to this rule are guanosine-containing nucleotides, which have a small preference for the *syn* form because of favorable electrostatic interactions between the exocyclic N2 amino group of guanine and the 5' phosphate atom. For pyrimidine nucleotides, the *anti* conformation is preferred over the *syn*, because of unfavorable contacts between the O2 oxygen atom of the base and the 5'-phosphate group. The results of molecular-mechanics energy minimizations on all four DNA nucleotides in both *syn* and *anti* forms (using the AMBER all-atom force field), are fully in accord with these observations (Table 2.2).

The sterically preferred ranges for the two domains of glycosidic angles are:

Anti: $-120 > \chi > 180^\circ$

Syn: $0 < \chi < 90^\circ$

Values of χ in the region of about -90° are often described as “high *anti*”. There are pronounced correlations between sugar pucker and glycosidic angle, which reflect the changes in non-bonded clashes produced by C2'-*endo* versus C3'-*endo* puckers. Thus, *syn* glycosidic angles are not found with C3'-*endo* puckers due to steric clashes between the base and the H3' atom, which points toward the base in this pucker mode.

2.6 The Backbone Torsion Angles and Correlated Flexibility

The phosphodiester backbone of an oligonucleotide has six variable torsion angles (Fig. 2.14), designated $\alpha \dots \zeta$, in addition to the five internal sugar torsions $\tau_0 - \tau_4$ and the glycosidic angle χ . As will be seen, a number of these have highly correlated values (and therefore correlated motions in a solution environment). Steric considerations alone dictate that the backbone angles are restricted to discrete ranges (Sundaralingam, 1969; Olson, 1982b) (Fig. 2.15), and are accordingly not free to adopt any value between 0° and 360° . Fig. 2.15 uses a conformational wheel to show these preferred values, which are directly readable from their positions around the wheel. The fact that angles α , β , γ , and ζ each have three allowed ranges, together with the broad range for angle ϵ that includes two staggered regions, leads to a large number of possible low-energy conformations for the unit nucleotide, especially when glycosidic angle and sugar pucker flexibility is taken into account. In reality, only a small number of DNA oligonucleotide and polynucleotide structural classes have actually been observed out of this large range of possibilities; this is doubtless in large part due to the restraints imposed by Watson–Crick base pairing on the backbone conformations when two DNA strands are intertwined together.

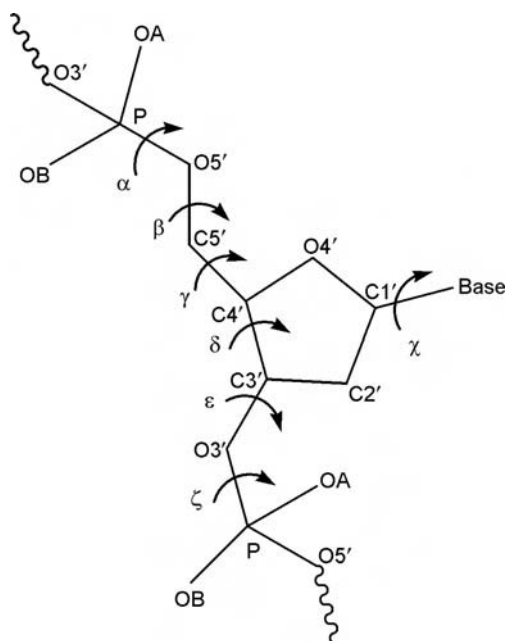


Figure 2.14 The backbone torsion angles in a unit nucleotide. Each rotatable bond is indicated by a curved arrow.

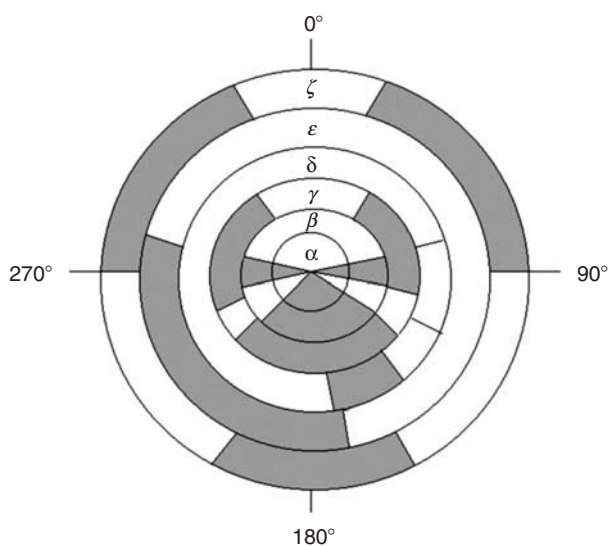


Figure 2.15 Conformational wheel showing the allowed ranges of backbone torsion angles (shaded) in nucleosides, nucleotides, deoxyoligonucleotides, and deoxypolynucleotides.

In contrast, crystallographic and NMR studies on a large number of standard and modified mononucleosides and nucleotides have shown their considerably greater conformational diversity, in accord with the possibilities indicated in Fig. 2.15. For these, backbone conformations in the solid state and in solution are not always in agreement; the requirements for efficient packing in the crystal can often overcome

the modest energy barriers between different values for a particular torsion angle. Many large RNA molecules are characterized by a wide range of base–base interactions, which can therefore adopt a wide variety of backbone conformations (Schneider, Moravsek, and Berman, 2004) (see Chap. 6 for a detailed discussion of RNA conformational variability).

A common convention for describing these backbone angles is to term values of $\sim 60^\circ$ as *gauche*⁺ (g^+), -60° as *gauche*⁻ (g^-) and $\sim 180^\circ$ as *trans* (t). Thus, for example, angles α (about the P–O5' bond) and γ (the exocyclic angle about the C4'–C5' bond), can be in the g^+ , g^- , or t conformations. The two torsion angles around the phosphate group itself, α and ζ , have been found to show a high degree of flexibility in various dinucleoside crystal structures, with the tg^- , g^-g^- , and g^+g^+ conformations all having been observed (Kim *et al.*, 1973). As will be described in Chap. 3, only the g^-g^- conformation can place successive nucleotide units in arrangements that have their bases in potential hydrogen-bonding positions with respect to a second nucleotide strand. Thus, this is the phosphate conformation for DNA and RNA double helices. The torsion angle β , about the O5'–C5' bond, is almost always found to be *trans*. All three possibilities for the γ angle have been observed in nucleoside crystal structures, although the g^+ conformation predominates in right-handed oligo- and polynucleotide double helices. The *trans* conformation for γ places the 5'-phosphate group in quite a distinct position with respect to the deoxyribose ring. The torsion angle δ around the C4'–C3' bond adopts values that relate to the pucker of the sugar ring, since the internal ring torsion angle τ_3 , (also around this bond), has a value of about 35° for C2'-*endo* and about 40° for C3'-*endo* puckers. δ is about 75° for C3'-*endo* and about 150° for C2'-*endo* puckers.

An alternative nomenclature for torsion angle ranges used by some workers in the nucleic acid field is that common in organic chemistry (the Klyne-Prelog system). In this, the *syn*(s) designation is given to angles clustered around 0° , and *anti* (a) for those around 180° . Intermediate angles are defined as $\pm$ *synclinal* ($\pm sc$) for around $\pm 60^\circ$, and $\pm$ *anticlinal* ($\pm ac$) for around $\pm 120^\circ$.

There are a number of well-established correlations involving pairs of these backbone torsion angles, as well as sugar pucker and glycosidic angle. These have been found (Sundaralingam, 1969) in mononucleosides and nucleotides (which are inherently more flexible in solution as well as being more subject to packing forces in the crystal), and more recently, in oligonucleotides (Schneider, Neidle, and Berman, 1997; Packer and Hunter, 1998; Varnai *et al.*, 2002). The existence of such correlations is important: it means that the atomic motions in oligo- and polynucleotides follow concerted patterns of inter-dependence. In general, these correlations are due to the diminution of non-bonded contacts that occur with particular conformations. Some of the more important correlations that have been observed in mononucleosides and nucleotides are:

- Between sugar pucker and glycosidic angle χ , especially for pyrimidine nucleosides. C3'-*endo* pucker is usually associated with median-value *anti* glycosidic angles, whereas C2'-*endo* puckers are commonly found with high *anti* χ angles. *Syn* glycosidic angle conformations show a marked preference for C2'-*endo* sugar puckers.
- The C4'–C5' torsion angle γ is correlated with the glycosidic angle and to some extent with sugar pucker and backbone angle α (Schneider, Neidle, and Berman, 1997; Varnai *et al.*, 2002). *Anti* glycosidic angles tend to correlate with g^+ conformations for γ .

References

- (1989). Definitions and nomenclature of nucleic acid structure parameters. *EMBO J.*, **8**, 1–4.
- Altona, C. and Sundaralingam, M. (1972). Conformational analysis of the sugar ring in nucleosides and nucleotides. A new description using the concept of pseudorotation. *J. Amer. Chem. Soc.*, **94**, 8205–8212.
- Babcock, M. S., Pednault, E. P. D., and Olson, W. K. (1994). Nucleic acid structure analysis. Mathematics for local Cartesian and helical structure parameters that are truly comparable between structures. *J. Mol. Biol.*, **237**, 125–156.
- Clowney, L., Jain, S. C., Srinivasan, A. R., Westbrook, J., Olson, W. K. and Berman, H.M. (1996). Geometric parameters in nucleic acids: nitrogenous bases. *J. Amer. Chem. Soc.*, **118**, 509–518.
- Dickerson, R. E. (1998). DNA bending: the prevalence of kinkiness and the virtues of normality. *Nucleic Acids Res.*, **26**, 1906–1926.
- Foloppe, N., Nilsson, L., and MacKerell, A. B., Jr. (2001). Ab initio conformational analysis of nucleic acid components: intrinsic energetic contributions to nucleic acid structure and dynamics. *Biopolymers*, **61**, 61–76.
- Gelbin, A., Schneider, B., Clowney, L., Hsieh, S.-H., Olson, W. K., and Berman, H. M. (1996). Geometric parameters in nucleic acids: sugar and phosphate constituents. *J. Amer. Chem. Soc.*, **118**, 519–529.
- Goodsell, D. S. and Dickerson, R. E. (1994). Bending and curvature calculations in B-DNA. *Nucleic Acids Res.*, **22**, 5497–5503.
- Gould, I. R. and Kollman, P. A. (1994). Theoretical investigation of the hydrogen bond strengths in guanine-cytosine and adenine-thymine base pairs. *J. Amer. Chem. Soc.*, **116**, 2493–2499.
- Harvey, S. C. and Prabhakaran, M. (1986). Ribose puckering: structure, dynamics, energetics, and the pseudorotation cycle. *J. Amer. Chem. Soc.*, **108**, 6128–6136.
- Kim, S.-H., Berman, H. M., Seeman, N. C., and Newton, M. D. (1973). Seven basic conformations of nucleic acid structural units. *Acta Crystallogr.*, **B29**, 703–710.
- Lavery, R. and Sklenar, H. (1988). The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids. *J. Biomol. Struct. Dynamics*, **6**, 63–91.
- Lavery, R. and Sklenar, H. (1989). Defining the structure of irregular nucleic acids: conventions and principles. *J. Biomol. Struct. Dynamics*, **6**, 655–667.
- Liu, D., Wyttenbach, T., and Bowers, M. T. (2006). Hydration of mononucleotides. *J. Amer. Chem. Soc.*, **128**, 15155–15163.
- Lu, X.-J. and Olson, W. K. (1999). Resolving the discrepancies among nucleic acid conformational analyses. *J. Mol. Biol.*, **285**, 1563–1575.
- Lu, X.-J., Babcock, M. S., and Olson, W. K. (1999). Overview of nucleic acid analysis programs. *J. Biomol. Struct. Dynamics*, **16**, 833–843.
- Lu, X.-J., El Hassan, M. A., and Hunter, C. A. (1997). Structure and conformation of helical nucleic acids: analysis program (SCHNAAP). *J. Mol. Biol.*, **273**, 668–680.
- Murray-Rust, P. and Motherwell, S. (1978). Computer retrieval and analysis of molecular geometry. III. Geometry of the β -1'-aminofuranoside fragment. *Acta Crystallogr.*, **B34**, 2534–2546.
- Olson, W. K. (1982a). How flexible is the furanose ring? 2. An updated potential energy estimate. *J. Amer. Chem. Soc.*, **104**, 278–286.
- Olson, W. K. (1982b) In *Topics in Nucleic Acid Structure*, Part 2 (Ed. Neidle, S.), pp. 1–79. Macmillan Press, London.
- Olson, W. K. and Sussman, J. L. (1982). How flexible is the furanose ring? 1. A comparison of experimental and theoretical studies. *J. Amer. Chem. Soc.*, **104**, 270–278.
- Olson, W. K., Bansal, M., Burley, S. K., Dickerson, R. E., Gerstein, M., Harvey, S. C., Heinemann, U., Lu, X.-J., Neidle, S., Shakked, Z., Sklenar, H., Suzuki, M., Tung, C.-S., Westhof, E., Wolberger, C., and Berman, H. M. (2001). A standard reference frame for the description of nucleic acid base-pair geometry. *J. Mol. Biol.*, **313**, 229–237.
- Packer, M. J. and Hunter, C. A. (1998). Sequence-dependent DNA structure: the role of the sugar-phosphate backbone. *J. Mol. Biol.*, **280**, 407–420.
- Parkinson, G., Vojtechovsky, J., Clowney, L., Brünger, A. T., and Berman, H. M. (1996). New parameters for the refinement of nucleic acid-containing structures. *Acta Crystallogr.*, **D52**, 57–64.
- Pearlman, D. A. and Kim, S.-H. (1990). Atomic charges for DNA constituents derived from single-crystal X-ray diffraction data. *J. Mol. Biol.*, **211**, 171–187.
- Rosenberg, J. M., Seeman, N. C., Day, R. O. and Rich, A. (1976). RNA double-helical fragments at atomic resolution. II. The crystal structure of sodium guanylyl-3',5'-cytidine nonahydrate. *J. Mol. Biol.*, **104**, 145–167.
- Schneider, B., Moravsek, Z., and Berman, H. M. (2004). RNA conformational classes. *Nucleic Acids Res.*, **32**, 1666–1677.

- Schneider, B., Neidle, S., and Berman, H. M. (1997). Conformations of the sugar-phosphate backbone in helical DNA crystal structures. *Biopolymers*, **42**, 113–124.
- Seeman, N. C., Rosenberg, J. M., Suddath, F. L., Kim, J. P., and Rich, A. (1976). RNA double-helical fragments at atomic resolution. I. The crystal and molecular structure of sodium adenylyl-3',5'-uridine hexahydrate. *J. Mol. Biol.*, **104**, 109–144.
- Sponer, J., Jurecka, P., and Hobza, P. (2004). Accurate interaction energies of hydrogen-bonded nucleic acid base pairs. *J. Amer. Chem. Soc.*, **126**, 10142–10151.
- Stofer, E., Chipot, C., and Lavery, R. (1999). Free energy calculations of Watson–Crick base pairing in aqueous solution. *J. Amer. Chem. Soc.*, **121**, 9503–9508.
- Sun, G., Voigt, J. H., Filippov, I. V., Marquez, V. E., and Nicklaus, M. C. (2004). PROSIT: Pseudo-rotational online service and interactive tool, applied to a conformational survey of nucleosides and nucleotides. *J. Chem. Inf. Comput. Sci.*, **44**, 1752–1762.
- Sundaralingam, M. (1969). Stereochemistry of nucleic acids and their constituents. IV. Allowed and preferred conformations of nucleosides, nucleoside mono-, di-, tri-, tetraphosphates, nucleic acids and polynucleotides. *Biopolymers*, **7**, 821–860.
- Varnai, P., Djuranovic, D., Lavery, R., and Hartmann, B. (2002). Alpha/gamma transitions in the B-DNA backbone. *Nucleic Acids Res.*, **30**, 5398–5406.
- Wain-Hobson, S. (2006). The third bond. *Nature*, **439**, 539.
- Watson, J. D. and Crick, F. H. C. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, **171**, 737–738.

Further Reading

- Jeffrey, G.A. and Saenger, W. (1991). *Hydrogen Bonding in Biological Structures*. Chaps. 15, 16. Springer, Berlin.
- Sponer, J., Leszczynski, J., and Hobza, P. (2001). Electronic properties, hydrogen bonding, stacking, and cation binding of DNA and RNA bases. *Biopolymers*, **61**, 3–31.
- Voet, D. and Rich, A. (1970). The crystal structures of purines, pyrimidines and their intermolecular complexes. *Prog. Nucleic Acid Res. Mol. Biol.*, **10**, 183–265.

DNA Structure as Observed in Fibers and Crystals

3.1 Structural Fundamentals

3.1.1 Helical Parameters

The fiber diffraction method for determining the structures of polymeric DNA (and RNA) molecules has been outlined in Chap. 1. Examination of X-ray diffraction photographs of oriented fibers obtained from them enables the basic parameters defining the dimensions of the repeating helix, to be calculated. Measurement of the spacing between the layer lines directly gives the reciprocal of the helical pitch P , defined as the distance, parallel to the helix axis, between successive nucleotide units per complete turn of the helix (Fig. 3.1). The presence of a meridional reflection for a fiber that is oriented perpendicular to the direction of the X-ray beam, indicates the presence of a structural periodicity in that direction. Analysis of the meridional reflection gives the regular repeat distance d (the helical rise if the nucleotide units are perpendicular to the helix axis). The unit repeat n , the number of nucleotide units in a single full turn of helix, that is per complete pitch P , is thus P/d . Helical rise is more generally the distance between successive nucleotide units, projected to be parallel to the helix axis if the units are not strictly perpendicular to this axis. The diameter of the helix can be derived from the dimensions of the closely packed unit cell projection down the fiber axis.

3.1.2 Base-Pair Morphological Features

Base pairs are not necessarily planar, and indeed are rarely so. This is in part because the geometric requirements of hydrogen-bonding are not especially stringent, and in particular the angle subtended at the hydrogen atom (donor-H-acceptor) can deviate by up to about 35° from linearity without appreciable loss of hydrogen-bond energy. Other factors such as the need to avoid steric clash in some base-pair-base-pair nonbonded interactions within a helix, can result in distortions of base pairs from planarity, since these do not distort hydrogen bonds beyond these energetically favorable limits. These deviations from planarity can take place in a number of ways that can

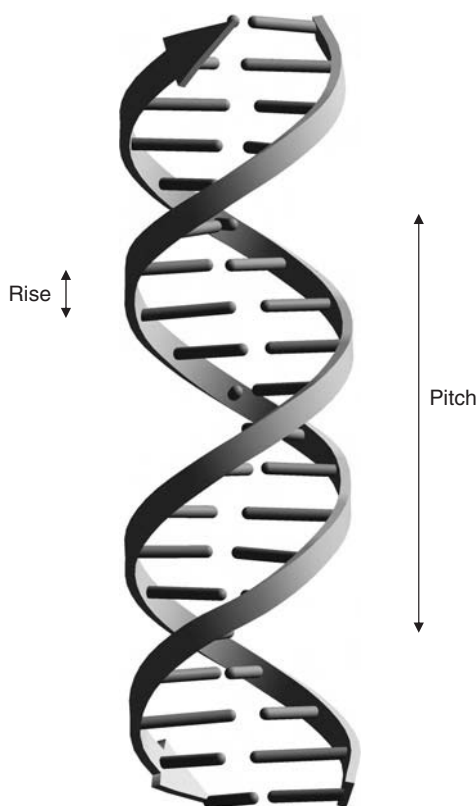


Figure 3.1 Schematic DNA double helix, with the backbones shown in ribbon representation.

be categorized into two groups of local base-pair morphological features (Dickerson *et al.*, 1989; Olson *et al.*, 2001), as described in Chap. 2. The first category involves individual base pairs. It specifies both movements of one base relative to another within a base pair (e.g., propeller twist and buckle), and movement of the base pair relative to global axes in the helix. The second category is concerned with the relative movements of base pairs in successive base pairs (base-pair steps), for example, helical twist and roll. It defines the relative relationships of successive base pairs in a structure. Inevitably the second set of features incorporate the first, so correlations between them can be expected.

3.2 Polynucleotide Structures from Fiber Diffraction Studies

3.2.1 Classic DNA Structures

Information on the dimensions of polynucleotide double helices was initially derived from the diffraction patterns of DNA fibers at high (92%) relative humidity, which showed a readily interpretable and characteristic Maltese cross pattern of strong diffraction intensities. Examination of this pattern, which arises from the B-DNA double helix, led directly to the original Watson–Crick model in 1953 (Watson and

Crick, 1953). Subjecting DNA fibers to other conditions produce rather different diffraction patterns. In particular, lower (65–75%) relative humidity conditions result in the A-DNA pattern, which typically has many more diffraction maxima, indicating greater crystallinity and hence order in the fibers. A number of other forms have subsequently been found (see Sect. 3.2.2), most of which are subclasses of the A and B double helices, the two most important ones for random sequence DNA. Their exactly repetitious structures are often referred to as “canonical DNA” ones. Double-helical DNA is thus highly polymorphic, the differing forms corresponding to distinct yet interconvertible molecular structures. Some DNA structural types have only been found with defined-sequence polynucleotides rather than with random-sequence DNA (the standard A and B forms can occur with most sequences). Under appropriate experimental conditions, the diffraction pattern of such a polynucleotide is best fitted to a repeating unit, that rather than being a simple mononucleotide, can be a dinucleotide or even an oligonucleotide. Some of these repetitious fibers produce semicrystalline diffraction patterns that give intensity data to significantly higher resolution than classic calf thymus DNA fibers. In these instances, the relatively high number of observed diffraction intensities, can result in structures that are as reliable as those of standard A and B polynucleotide helices.

B-DNA, the classic structure first described by Watson and Crick, has subsequently been refined (Arnott, 1999) using the linked-atom least-squares procedure developed by Arnott and his colleagues. The helical repeat is a single nucleotide unit, with necessarily all of the nucleotides in the structure having the conformation of this unit. Helical parameters for B- and other DNA polynucleotides are given in Table 3.1. The backbone conformation (Table 3.2) has high *anti* glycosidic angles and C2'-*endo* sugar puckers (Fig. 3.2).

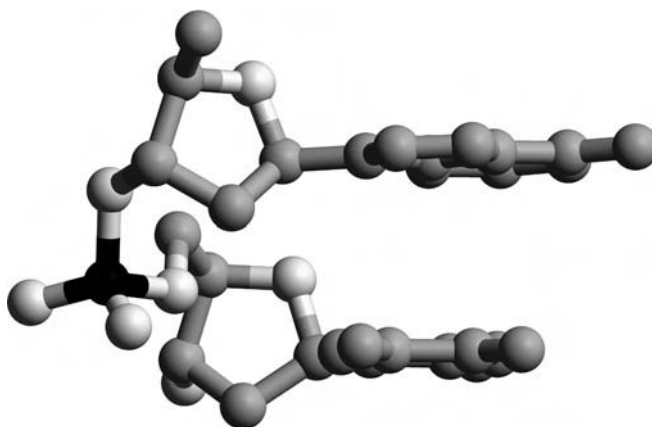
The right-handed double helix has 10 base pairs per complete turn, with the two polynucleotide chains wound antiparallel to each other (Fig. 3.3a) and linked by Watson-Crick A•T and G•C base pairs. The paired bases are almost exactly perpendicular to the helix axis, and they are stacked over the axis itself (Fig. 3.3b). Consequently the base-pair separation is the same as the helical rise – 3.4 Å. An important consequence of the Watson-Crick base-pairing arrangement is that the two deoxyribose sugars linked to an individual base pair are on the same side of it. So, when successive base pairs are stacked on each other in the helix, the gap between these sugars forms continuous indentations in the surface that wind along, parallel to the sugar-phosphodiester chains. These indentations are termed grooves. Groove width can be defined as the perpendicular distance between phosphate groups on opposite strands,

Table 3.1 Selected Helical Parameters for Various Polymorphs of DNA Polynucleotides, Derived from Fiber Diffraction Studies. (Values are Mostly from Arnott, 1999)

	Unit repeat	Rise (Å)	Helical twist (°)	Base pair propeller twist (°)	Base step roll (°)	Base pair inclination (°)
A	11	2.54	32.7	–10.5	0.0	22.6
B	10	3.38	36.0	–15.1	0.0	2.8
C	28/3	3.31	38.6	–1.8	0.0	–8.2
D	8	3.01	45.0	–21.0	0.0	–13.0
Z(C)	6	7.25	–49.3	8.3	5.6	0.1
Z (G)	6	7.25	–10.3	8.3	–5.6	0.1

Table 3.2 Backbone Conformational Angles in (°) for Various Polymorphs of DNA Polynucleotides. (Taken from Arnott, 1999)

	α	β	γ	δ	ϵ	ζ	χ
A	-52	175	42	79	-148	-75	-157
B	-30	136	31	143	-141	-161	-98
C	-37	-160	37	157	161	-106	-97
D	-59	156	64	145	-163	-131	-102
Z (C)	-140	-137	51	138	-97	82	-154
Z (G)	52	179	-174	95	-104	-65	59

**Figure 3.2** Conformation of two successive nucleotides in one strand of canonical B-DNA, from fiber diffraction analysis.

minus the van der Waals diameter of a phosphate group (5.8 Å), although it is equally valid in some circumstances to use pairs of other atoms, such as C1' or O4'. Groove depths are normally defined in terms of the differences in cylindrical polar radii between phosphorus and N2 guanine or N6 adenine atoms, for minor and major grooves respectively.

The asymmetry in the base pairs results in two parallel types of grooves, whose dimensions (especially their depths) are related to the distances of base pairs from the axis of the helix and their orientation with respect to the axis. The B-DNA wide major groove (Fig. 3.3c and Table 3.3) is almost identical in depth to the much narrower minor groove, which has the hydrophobic hydrogen atoms of the sugar groups forming its walls. In general, the major groove is richer in base substituents – O6, N6 of purines and N4, O4 of pyrimidines compared to the minor one. This, together with the steric differences between the two, has important consequences for interaction with other molecules (see Chaps. 4–7). Base-pair and base-step morphologies for several polynucleotide helices are given in Table 3.1.

The A-DNA duplex also has a single nucleotide unit as the helical repeat (Fig. 3.4). This has C3'-*endo* sugar puckers, which bring consecutive phosphate groups on the nucleotide chains closer together – 5.9 Å compared to the 7.0 Å in B-form DNA, and alters the glycosidic angle from high *anti* to *anti*. As a consequence, the base pairs are

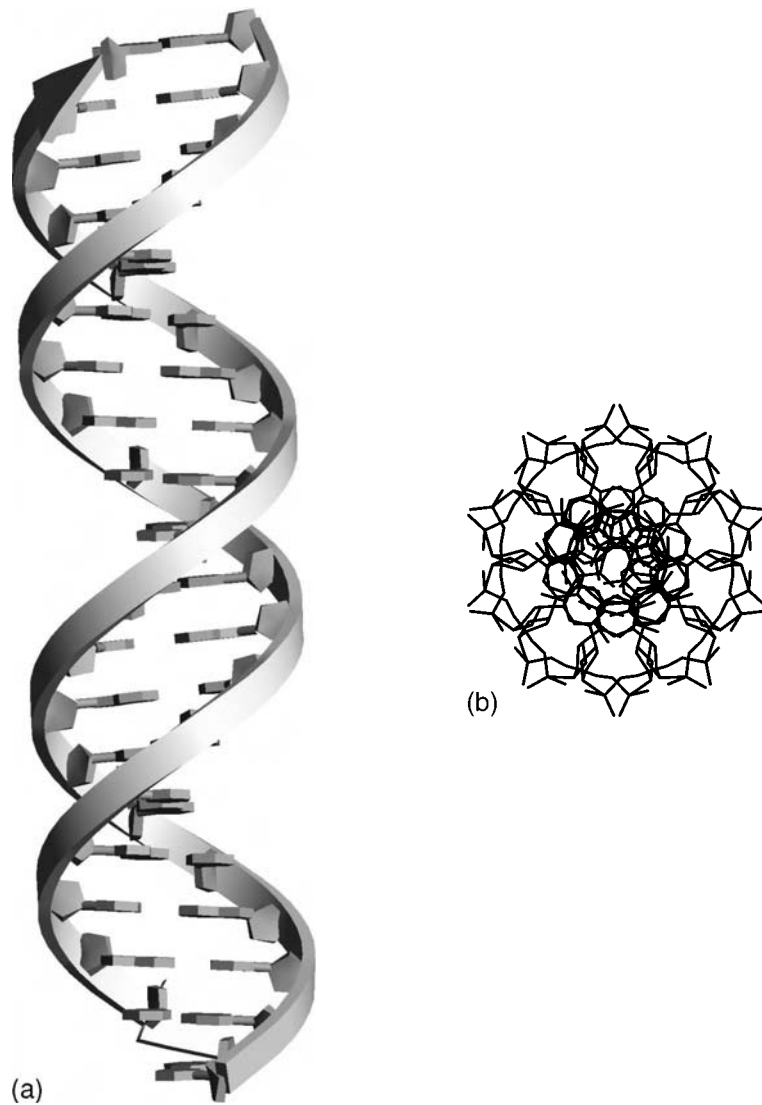


Figure 3.3 (a) The structure of canonical B-DNA, from fiber diffraction analysis; (b) View down the helix axis;

twisted and tilted with respect to the helix axis (Fig. 3.5a), and are displaced nearly $5 \text{ \AA}$ from it, in striking contrast to the B helix. The helical rise is as a consequence much reduced, to $2.54 \text{ \AA}$, compared to $3.4 \text{ \AA}$ for canonical B-DNA. The helix is wider than the B one and has an 11 base-pair helical repeat. The combination of base-pair tilt with respect to the helix axis and base-pair displacement from the axis results in very different groove characteristics for the A double helix compared to the B form (Fig. 3.5b). This also results in the center of the A double helix being a hollow cylinder (Fig. 3.5c). The major groove is now deep and narrow, and the minor one is wide and very shallow.

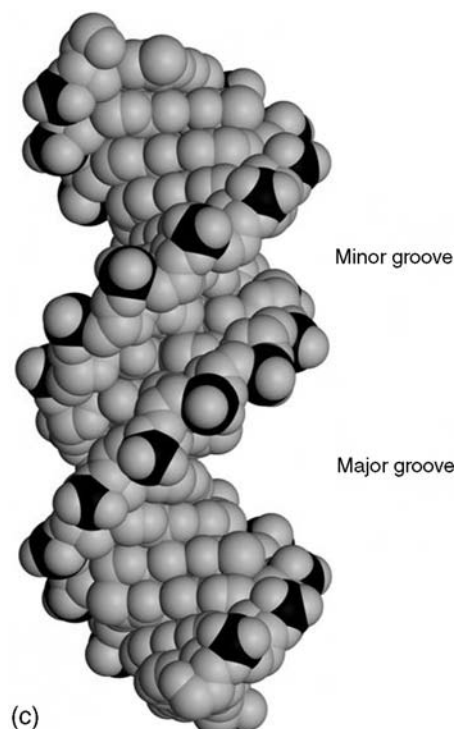


Figure 3.3 Cont'd. (c) Space-filling representation of B-DNA, showing major and minor grooves. Phosphorus atoms are colored black.

Table 3.3 Groove Dimensions (in Å) in Polynucleotide DNA Double Helices, from Fiber Diffraction Analyses. (from Arnott, 1999 for the A – D Helices and from the Single-Crystal Data (Wang *et al.*, 1979, 1981) for the Z Helix)

Polymorph	Major groove		Minor groove	
	Width	Depth	Width	Depth
A	2.2	13.0	11.1	2.6
B	11.6	8.5	6.0	8.2
C	10.5	7.6	4.8	7.9
D	9.6	6.2	0.8	7.4
Z	8.8	3.7	2.0	13.8

3.2.2 DNA Polymorphism in Fibers

The structural transition from A to B forms in “mixed-sequence” DNA is induced by changes in the relative humidity surrounding the DNA fibers. Other forms can be produced by fine control of this condition (the C form) or from defined-sequence polynucleotides (the D and Z forms). For example, there are a number of members of both the A and B families that differ from the “canonical” forms outlined above, in large part because of the degree of overwinding of the helices (for the B-type family), and have considerable differences in pitch, base-pair tilt with respect to the helix axis, and hence groove characteristics. There are variants on the A-DNA duplex that have a wide major

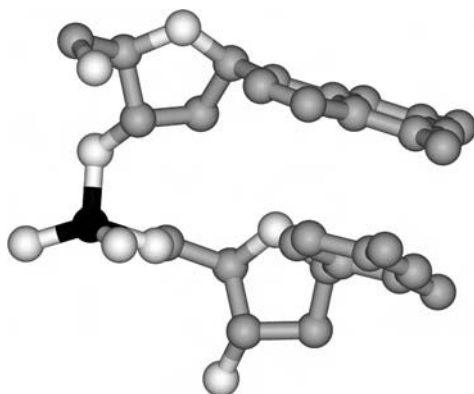


Figure 3.4 Conformation of two successive nucleotides in one strand of canonical A-DNA, from fiber diffraction analysis.

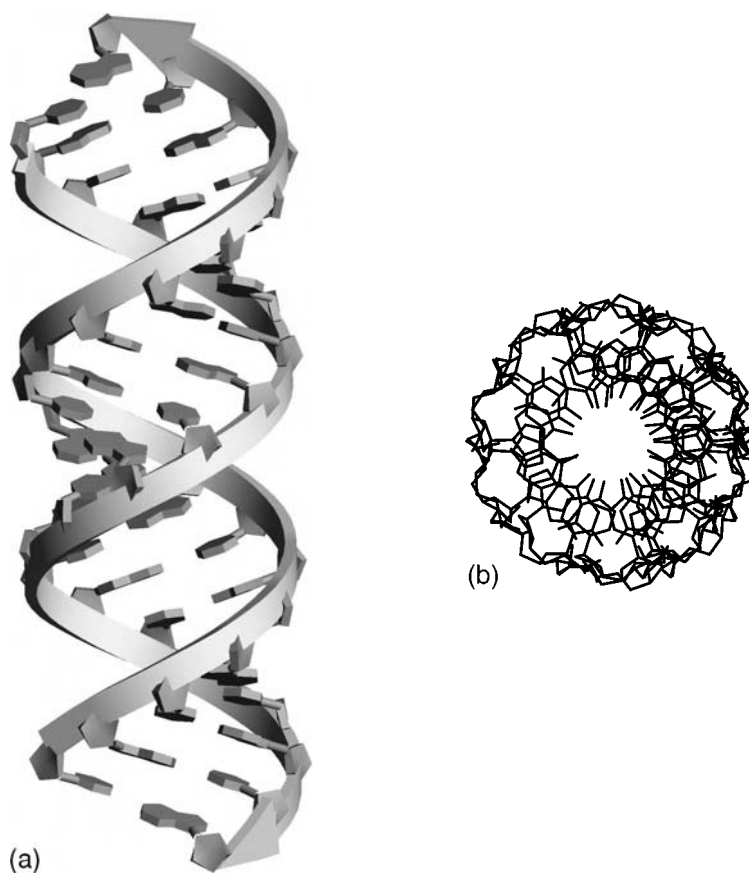


Figure 3.5 (a) The structure of canonical A-DNA, from fiber diffraction analysis, in ribbon representation; (b) View down the helix axis;

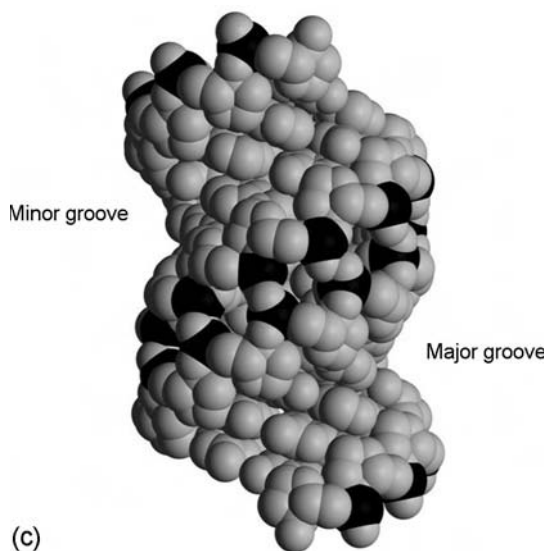


Figure 3.5 Cont'd. (c) Space-filling representation of A-DNA, showing major and minor grooves.

groove (Arnott *et al.*, 1982), reminiscent of that in a B helix, rather than that of the narrow standard A groove. Changes in groove width can thus be achieved at little energy cost.

The C form of DNA results from relatively low-humidity conditions for a DNA fiber and is overwound relative to B-DNA, with 9.3 residues per turn. The overall appearance of this helix and the dimensions of the grooves resembles B-type DNA rather than the A-form morphologies. The D form cannot be adopted by random sequence native DNAs, and has been observed in crystalline fibers of the alternating purine/pyrimidine polynucleotides poly(dA-dT)•poly(dA-dT) and poly(dI-dT)•poly(dI-dT), where I is inosine (the rare 8-oxo-purine). The structure of this polymorph has not been unambiguously defined as yet, with models proposed ranging from left-handed seven- and eightfold helices to a more conventional eightfold right-handed one. Structural parameters for this right-handed model are given in Tables 3.1–3.3. Observations of the reversible structural transition from B to D forms in crystalline fibers have been made using time-resolved X-ray diffraction by means of a high-intensity synchrotron source of X-rays (Mahendrasingam *et al.*, 1986). The observation of a gradual rather than an abrupt increase in helical pitch, from 24 Å to 34 Å, is consistent only with a transition in which there is no change in helix-handedness, and so a left-handed D-DNA model can be rejected.

The Z form is an authenticated left-handed structure (Arnott *et al.*, 1980). It exists in the alternating sequence poly(dC-dG)•poly(dC-dG) and is presumed to be the structure formed by this sequence in solution in high salt (>2.5 M NaCl) conditions. This left-handed DNA has a dinucleotide repeat (Fig. 3.6) with quite distinct nucleoside conformations for the guanosine compared to the cytosine residues. Each individual repeat has a helical rise of 7.25 Å, so that the rise between successive base pairs is half of this – 3.63 Å (Fig. 3.7). Z-DNA is discussed further in Sect. 3.5 of this chapter; it was discovered in a defined-sequence oligonucleotide crystal structure as well as by fiber diffraction analysis of poly(dC-dG)•poly(dC-dG).

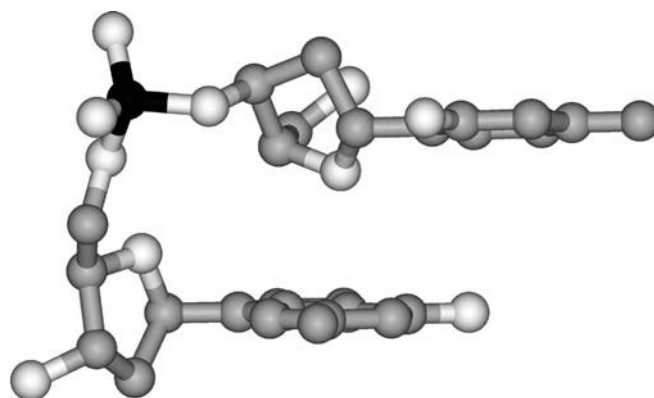


Figure 3.6 Conformation of two successive nucleotides in one strand of canonical Z-DNA, from fiber diffraction analysis.



Figure 3.7 The Z-DNA left-handed double helix, in ribbon representation to emphasize the discontinuities in the backbone.

The polymorphism of DNA structure that is apparent in polymeric fibers, is suggestive of an underlying flexibility in DNA structure. This is a natural consequence of the large number of backbone and sugar conformational variables, which together with base-pair flexibility, can result in double helices that are structurally distinct yet are equivalent in energetic terms. Fiber diffraction studies can only hint at the fine detail of this flexibility. So further significant advances in DNA structure studies have taken place with the advent of single-crystal and NMR analyses of defined oligonucleotide sequences, which as described in Chap. 1, provide structural data that is not averaged over the sequence. These studies have provided much detail not only of the structures themselves, but also are enabling the principles to be determined that relate DNA structural features to sequence and environment.

3.3 B-DNA Oligonucleotide Structure as Seen in Crystallographic Analyses

3.3.1 The Dickerson–Drew Dodecamer

The single-crystal structure of the self-complementary dodecanucleotide $d(\text{CGCGAATTCGCG})_2$ was determined in 1979 by multiple isomorphous replacement methods (Wing *et al.*, 1980). This structure analysis is of historic importance in that without recourse to any preconceived model, it showed an antiparallel right-handed B-DNA double helix (Fig. 3.8a), and thus unequivocally demonstrated the correctness of the Watson–Crick model for DNA structure. This, the “Dickerson–Drew” sequence has subsequently been widely studied by a variety of other experimental and theoretical techniques, that have complemented, generally confirmed, and sometimes extended the original structural results. The original crystallographic analysis, at 1.9 Å resolution, also revealed a number of major sequence-dependent structural features that could not be observed in fiber diffraction studies of averaged-sequence B-DNA, which has all the nucleotides in an identical conformation. These features are:

- A narrow minor groove in the 5′-AATT region (Fig. 3.8b), which at its extreme point is only 3.2 Å wide compared to 6.0 Å in averaged canonical B-DNA from fiber diffraction studies. The major groove at this point in the dodecamer structure is exceptionally wide (12.7 Å).
- A well-ordered and regular network of water molecules in this A/T region of the minor groove (Drew and Dickerson, 1981). This network has been termed the “spine of hydration.” It involves hydrogen-bonding of first-shell water molecules to the O2 atom of thymine and the N3 atom of adenine such that each water molecule spans the two dodecamer strands. These water molecules are linked together by further, second-shell waters. Evidence for the persistence of this structured water arrangement in solution has come from NMR studies (Kubinec and Wemmer, 1992), from high-resolution crystallographic studies (Vlieghe, Turkenberg, and Van Meervelt, 1999), and from a number of molecular dynamics simulation studies (see Rueda *et al.*, 2004 for a comprehensive survey of earlier work). Minor-groove hydration and the structure of the spine are discussed in detail in Chap. 5.
- Differences in the values for a number of local base pair and base-step parameters at different points along the sequence (Dickerson and Drew, 1981). Most notable are the high propeller twist values for the four central A•T base pairs (Fig. 3.9).

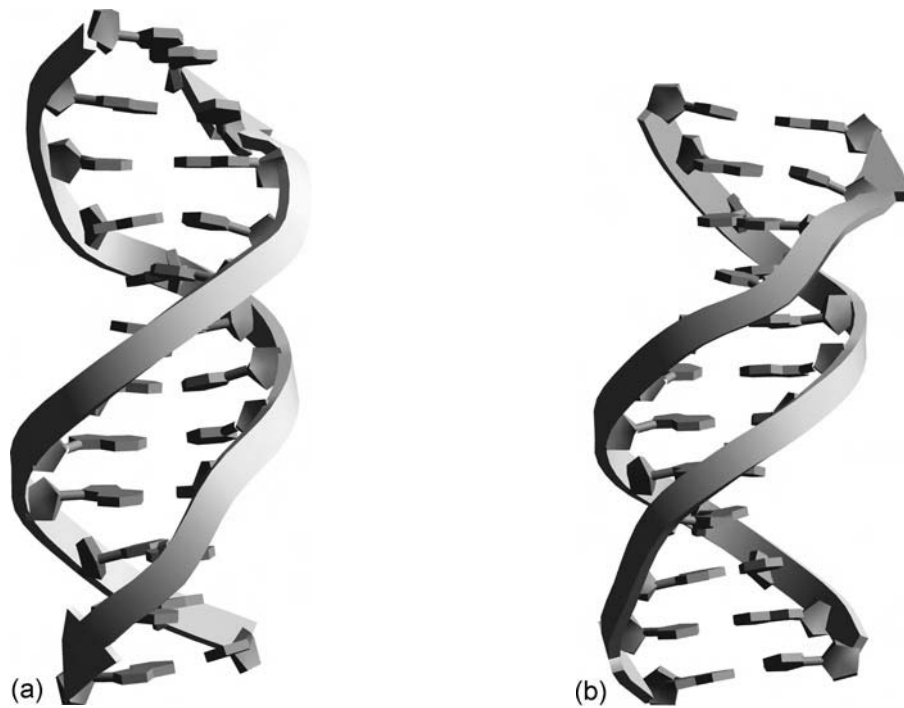


Figure 3.8 (a, b) Two ribbon views of the structure of the Dickerson–Drew dodecamer, from the first crystallographic analyses of this sequence (Wing *et al.*, 1980; Drew and Dickerson, 1981). The two views are rotated by $\sim 45^\circ$, and show the narrowing of the minor groove in the central region of the sequence as well as slight irregularities in the backbone conformation.



Figure 3.9 View of the central region in the Dickerson–Drew dodecamer, showing the high propeller twist of the A•T base pairs.

- A wide distribution of values for sugar puckers, glycosidic angles, and backbone conformational angles. The value for a particular torsion angle is to a considerable degree dependent on the sequence context of the nucleotide involved. For example, values for torsion angles ϵ and ζ are distributed in two subgroups within their broad ranges. Both angles are normally in the trans,

gauche⁻ domain, but are *gauche*⁻, *trans* when the nucleotide is followed by a purine. This alternative phosphate conformation has been designated B_{II}, with the standard *trans*, *gauche*⁻ phosphate conformation being termed B_I. The B_{II} conformation is associated with increased minor-groove width, both in the dodecamer and in a number of oligonucleotide crystal structures that have been subsequently determined. Its occurrence is mostly at the dinucleotide steps GpC, CpG, CpA, TpG, and TpA (Madhumalar and Bansal, 2005). Several other correlations between backbone torsion angles were found in the original Dickerson–Drew structure, and subsequently confirmed in a number of oligonucleotide structures, such as those between angles α and γ , and between angle δ and glycosidic angle χ . The range of values for the δ angle is itself sugar-pucker-dependent; this important correlation confirms and extends earlier findings from surveys of mononucleoside structures (Chap. 2, Sect. 2.5), to the oligonucleotide level.

- The average features of the structure are close to those for canonical B-DNA from fiber diffraction. For example, the average helical twist is 35.9° and there are 10.1 base pairs per helical turn.
- The helix is not straight, but is bent by about 19° in the major-groove direction. This bending is apparent in Fig. 3.8b.

3.3.2 Other Studies of the Dickerson–Drew Dodecamer

The role played by lattice interactions in the dodecanucleotide crystal structure has been examined in detail by several studies. Comparison of nine different dodecamers in three space groups has shown that the key parameters of minor-groove width and propeller twist for the central 5'-AATT region are largely unaffected by crystal-packing factors (Dickerson, Goodsell, and Neidle, 1994). This is to be expected since in the P2₁2₁2₁ original space group for the Dickerson–Drew (and a number of other) dodecamers, only the terminal two nucleotides at each end are involved in intermolecular interactions with other duplexes. Thus the central helical region of 6–8 base pairs is unconstrained, and its features are intrinsic to the particular sequence. In a more demanding test of the generality of the features it has been found that d(CGCGAATTCGCG) can crystallize in an alternative space group (the trigonal R3) with a distinct packing arrangement, when under the influence of particular metal ions such as calcium (Liu and Subirana, 1999; Minasov, Tereshko, and Egli, 1999).

In all cases the key features of base and base-pair morphology are still apparent, although sequence-dependent features tend to differ less from canonical values. These together with the narrow minor groove filled by a spine of hydration, provide good evidence for the general correctness of the original structure. All of these structures contain both dodecanucleotide strands in the crystallographic asymmetric unit, and the resulting duplex is not perfectly twofold symmetric, by contrast with its structure in solution. A further crystal form (Johansson, Parkinson, and Neidle, 2000) in the trigonal space group P3₂12, by contrast, has the duplex sitting on a crystallographic twofold axis, so that exact twofold symmetry is imposed on the structure (Fig. 3.10). Features such as the narrow minor groove and hydration spine are still apparent, but now they are twofold symmetric, reflecting the exact symmetry implied by the self-complementary Dickerson–Drew sequence.



Figure 3.10 The structure of the dodecamer d(CGCGAATTCGCG) in space group $P3_12$, where the duplex utilizes a two fold crystallographic symmetry axis (Johansson, Parkinson, and Neidle, 2000). The terminal nucleotides are not base-paired within the duplex; instead they are hydrogen-bonded to other duplex molecules in the crystal lattice.

The status of the Dickerson–Drew structure as the paradigm for many of the fine details of B-DNA structure generally, implies that its features are considered to be totally reliable, especially by nonexperts in crystallography. It is important to bear in mind that the original structure was determined to a resolution that is only moderate by current standards, and refined by methods that are now considered to be suboptimal. Thus the accuracy and precision of this analysis are lower than more recent studies on this and other sequences. These have used (i) high-quality, highly redundant diffraction data from synchrotron sources, (ii) libraries of high-precision nucleotide geometry, and (iii) refinement methods that are able to overcome false local minima. The reassuring picture that has emerged is one in which the central concept of sequence-dependent structure is retained, although the fine detail may differ from conclusions based on medium-resolution structures.

The crystal structures of the sodium and magnesium forms of the Dickerson–Drew sequence have been determined to 1.4 Å and 1.1 Å resolution respectively (Minasov, Tereshko, and Egli, 1999; Shui *et al.*, 1998; Tereshko, Minasov, and Egli, 1999). These structures confirm many of the above details, apart from the assignment of solvent in the minor groove, with the analysis being consistent with primary solvent sites in the spine of hydration-sharing between sodium cations and water molecules (the arrangement is described in detail in Chap. 5). Analysis of minor-groove width in the Dickerson–Drew structure using molecular dynamics simulations has led to the suggestion (Hamelberg *et al.*, 2000; Hamelberg, Williams, and Wilson, 2001) that it changes as a result of cation interactions, and is thus dependent on the

nature of the cation, as well as on sequence in the minor groove. The structure of the Dickerson–Drew sequence in solution has been examined in detail by NMR techniques (see, e.g., Wu *et al.*, 2003), which have tended to suggest that there is some conformational heterogeneity, consistent with the picture from the most high-resolution X-ray structures. This conclusion has been supported and extended by a combined NMR and low-angle X-ray analysis (Schwieters and Clore, 2007), which has provided reliable data on the amplitudes of motion for B_I/B_{II} phosphate motions, sugar pucker, and base morphology.

3.3.3 Other B-DNA Oligonucleotide Structures

The crystal structures of standard B-type oligonucleotides mostly have averaged features that closely approximate those of averaged-sequence fiber-diffraction B-DNA. Few are more than 12 base pairs in length. Almost all are self-complementary palindromic sequences. They fall into two principal categories. They are either dodecamers, the overwhelming majority of which are isomorphous to the Dickerson–Drew structure and therefore pack identically in the crystal, in the space group P2₁2₁2₁, or they are decamers, which crystallize in a variety of space groups, some of which give very high-resolution (0.74–1.4 Å) structures. The isomorphism of the dodecamers provides a convenient framework to analyze systematic changes in the central base pairs, whilst keeping invariant the 3′ and 5′ terminal sequences, which are involved in the crystal-packing arrangements. Several decamers crystallize in forms with the 10-mer duplexes packed end to end in the crystal, in a quasi-helical manner. The Dickerson–Drew-related sequence d(CGCAATTGCG) has been crystallized in no less than five different space groups, showing some differences in crystal packing (Valls *et al.*, 2004). These result in differences in the terminal base pairs, whilst the central CAATTG sequence is almost invariant structure in all five structures, providing good evidence that its features are intrinsic to the sequence (Fig. 3.11).

A large number of crystal structures with changes to the central GAATTC sequence of the Dickerson–Drew dodecamer have been examined. When the sequence is a purely alternating one, as in d(CGCATATATGCG) (Yoon *et al.*, 1988), the minor-groove width is equally narrow as in the Dickerson–Drew structure, although the narrowest points do not coincide. This 5′-ATATAT sequence shows a pronounced alternation of several of its base-pair and step parameters, especially roll; propeller twist values for the A•T base pairs are consistently lower than those in the Dickerson–Drew sequence.

The structure of the duplex formed by d(CGCAAAAAGCG) and its complementary sequence (Nelson *et al.*, 1987) is remarkable for several reasons. It is a model sequence for an A tract, which in natural DNA has anomalous structural and dynamic properties – it cannot be readily be reconstituted into nucleosomes, and is intrinsically bent when in phase with a helical repeat (see Sect. 3.6). The structure has high propeller twists for the A•T base pairs, of up to 26°, with a network of bifurcated three-center hydrogen bonds connecting them (Fig. 3.12). These bifurcated hydrogen bonds are in the major groove and involve atoms N6 of adenines and O4 of thymines. They have been suggested to be major factors in stabilizing the high propeller twists for the A•T base pairs, and thus in stiffening the A tract, at least in this structure. The roll, tilt, and slide values for the ApT steps in the structure are all close to 0°, with bending (toward the major groove) occurring at the ends of the A tract rather than within it (Table 3.4).

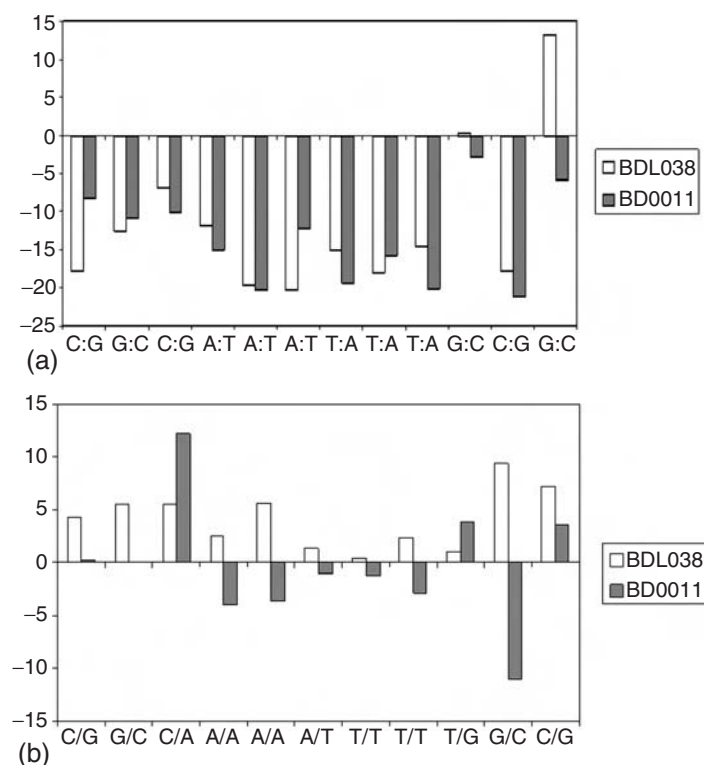


Figure 3.11 Plots of helical parameters in several dodecanucleotide crystal structures.

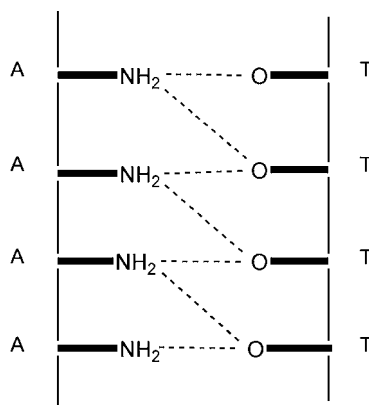


Figure 3.12 Schematic view of three-center major-groove bifurcated hydrogen-bonding involving A•T base pairs.

Bending in the major-groove direction has also been observed in the structure of the duplex formed by d(CGCAAAAATGCG) and its complementary strand (DiGabriele and Steitz, 1993), although the situation here is complicated by the presence of two distinct orientations of the duplex in the crystal lattice. This study has concluded that the bending found in the isomorphous dodecamer crystal structures is more a consequence of crystal-packing forces than of intrinsic properties of the A tract sequences.

Table 3.4 Selected B-DNA Crystal Structures. No NDB Accession Numbers are Available for those Structures that have Bound Protein. The 2nd Strand is not Listed for Sequences that are not Self-Complementary

Sequence	PDB number	NDB number
d(CGCGAATTCGCG) orthorhombic	1BNA	BLD0001
d(CGCGAATTCGCG) trigonal	1EHV	BD0032
d(CGCAAAAAAGCG)	1D98	BDL006
d(CGCAAAAATGCG)	1BDN	BDL015
d(CCGCTAGCGG)	1DCV	BD0028
d(CCAACGTTGG)	5DNB	BDJ019
d(CCAGTACTGG)	1D8G	BD0023
d(ACCGGCGCCACA)	330D	BDL035
d(CATGGGCCCATG)	1DC0	BD0026
d(CGCAAATTTGCG) at high resolution	1S2R	BD0067
d(CGCTGGAAATTTCCAGC)	1SGS	BD0070
d(CTTTTAAAGAAAAG)	1N4L	-
d(ATATAT)	1GQU	UD0035

The crystal structure of d(CGCAAATTTGCG) has been determined at medium (2.2 Å) (Edwards *et al.*, 1992) and high (1.4 Å) resolution (Woods *et al.*, 2004). Both structures show structured water molecules in the minor groove, although there are differences in groove width. These may reflect not only the superior resolution of the latter structure but also its improved data quality and the many improvements in refinement over a 12-year period. Even so, both structures concur in not finding significant three-center hydrogen-bonding in the A/T region, suggesting that these effects may not be inherently required to stabilize propeller twist and a narrow minor groove. Rather, propeller twist is more a consequence of base-stacking preferences.

The weight of evidence supports the concept of a narrow minor groove being preferred in A/T regions. However this is only a preference, and not a definitive rule obeyed under all circumstances. On the one hand, the duplex formed by d(CGCAAGCTGGCG) has an average minor-groove width of about 7 Å in the 5'-AGCT region (Webster *et al.*, 1990). This is close to that of averaged fiber diffraction B-DNA, as is the G/C-rich minor groove in the d(ACCGGCGCCACA) duplex (Timsit *et al.*, 1989). On the other hand, the high-resolution (1 Å) structures of the decamers d(CCAACGTTGG) and d(CCAGCGCTGG) both show a pronounced narrowing in the central region of the minor groove (Chiu and Dickerson, 2000), even though individual base-pair propeller twists are generally no greater than about 12°. The flexibility of such structures is further shown in the crystal structure (Kielkopf *et al.*, 2000) of the decamer d(CCAGTACTGG). This is notable in being one of the very few DNA duplex crystal structures to be determined at true atomic resolution (0.74 Å), so that the accuracy and precision of atomic coordinates and derived parameters is fully established. (The average esds of bond lengths and angles in this structure are 0.028 Å and 2.2° respectively). The structure shows that a number of regions of the backbone adopt more than one conformation, an effect that is undetected at lower resolution. These alternative conformations do not affect the general pattern of minor-groove narrowing that is centered around the short two base-pair A-tract. These A•T base pairs have high propeller twists of -20.3° and -18.8°, whereas the two adjacent C•G ones have values of -8.4° and -9.7° respectively.

The nonpalindromic sequence d(ACCGGCGCCACA) (Timsit *et al.*, 1989) is a hot-spot for frameshift mutagenesis by several chemical carcinogens. The structure is

nonisomorphous with the other dodecamers and is remarkable for several features. Three G•C base pairs in the center of the helix are partially opened, with exceptionally high propeller twists and some loss of G•C hydrogen-bonding. It has been suggested that this represents the intermediate stage prior to full base-pair opening. Analysis of a low temperature form of the crystals has shown (Timsit, Vilbois, and Moras, 1991) that cooling produces a one base-pair shift of the major-groove base-pairing that is favored by CA tracts. The crystal packing involves helices interacting together by means of groove-backbone interactions (Timsit and Moras, 1991), which are analogous to true DNA–DNA junctions such as the Holliday junction involved in recombination processes (see Chap. 4). This pattern contrasts with the packing observed in crystal structures of all-A/T oligonucleotides, most of which adopt a closely similar pattern of stacked parallel duplexes, often with staggered ends so that quasi-infinite helices are produced (Campos *et al.*, 2006). Such an arrangement also resembles the packing in polynucleotide fibers, and analogous orientational disorder may be observed, so that detailed single-crystal analyses are not always possible, as in the structure of d(ATATATATATAT), whose fiber-diffraction pattern has been interpreted in terms of a hexanucleotide-repeat coiled-coil structure with a pitch of 218.6 Å (Campos *et al.*, 2005). The crystal structure of one particular A/T sequence, that of d(ATATAT), is remarkable in showing an antiparallel duplex with Hoogsteen rather than Watson-Crick base pairs (Abrescia *et al.*, 2002). The right-handed double helix has overall similarity to the standard B-form in that the sugar puckers are C2'-*endo*, there are 10.6 base pairs per turn, and the base pairs are perpendicular to the helix axis (Fig. 3.13). The glycosidic angles for all the adenosine nucleotides are required to be in the *gauche*⁺ range for the *syn* conformation, as indeed is observed.

Crystallographic analyses of a number of the diverse sequences that are able to crystallize as decamer duplex helices (e.g., Quintana *et al.*, 1992; Lipanov *et al.*, 1993; Goodsell, Grzeskowiak, and Dickerson, 1995; Heinemann, Alings, and Bansal, 1992) has provided further evidence for variations in sequence-dependent structural features being real effects rather than artefacts. Comparisons of the structures show that the parameters helical twist, roll, and rise are not independent of each other, but that there are several correlations between them. For example, rise is linearly related to helical

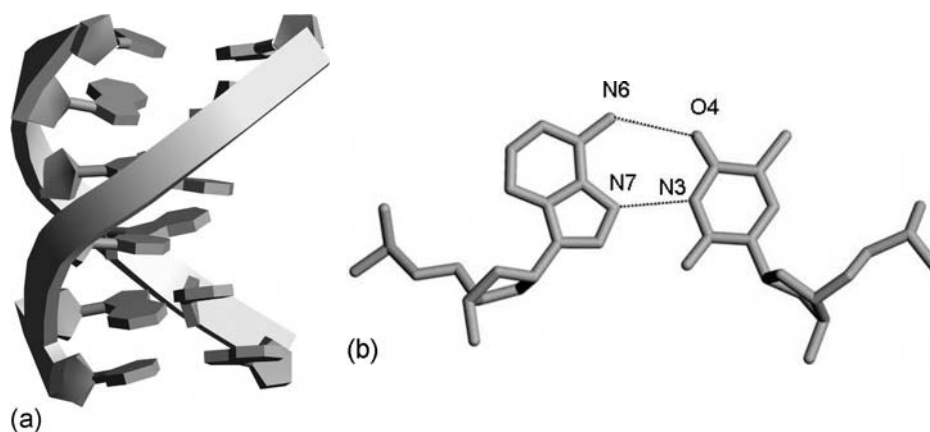


Figure 3.13 Two views of the d(ATATAT) structure (Abrescia *et al.*, 2002). (a) showing one of the two antiparallel duplexes and (b) one of the A•T Hoogsteen base pairs in the structure.

twist. The TpA step, even though it is consistently found to be unstacked in all the B-DNA oligonucleotide structures where it occurs, has highly variable sequence and crystal environment-dependent behavior. An extreme and surprising example of this is in the sequence d(CGATATATCG), which has a wide minor groove in the alternating AT region (Yuan, Quintana, and Dickerson, 1992), albeit with a characteristic alternation of helical twist angles that is a consequence of the alternation of good base-stacking at ApT steps combined with the poor stacking at the TpA ones. Thus, although alternating A/T sequences generally have a marked tendency to produce narrow minor grooves in B-DNA, their inherent structural flexibility means that this tendency can be readily overcome under the influence of TpA unstacking. The existence of the B_{II} phosphate conformation, which often occurs at dinucleotide steps in which the second base is a purine, is itself a factor in producing a widened minor groove (Hartmann, Piazzola, and Lavery, 1993). However the B_I/B_{II} backbone conformational flexibility observed in several high-resolution crystal structures (Kielkopf *et al.*, 2000; Schuerman and van Meervelt, 2000) suggests that this is less a factor than the inherent properties of the bases and base steps themselves.

Crystallographic studies using X-rays are generally not able to define the positions of hydrogen atoms in relatively mobile water molecules, even in the highest-resolution crystal structures. The methodology of neutron diffraction is able to do so since neutrons are strongly scattered by hydrogen (and deuterium) atoms. This technology has the disadvantages that high-flux neutron sources with diffraction facilities are only available in a few centers worldwide, and that large crystals are required, with a volume >1 mm³, which tends to limit the observable resolution. Crystals of up to 2.8 mm³ volume have been grown from D₂O solution (Arai *et al.*, 2002) for the sequence d(CCATTAAATGG), which has been previously studied by X-ray methods (Goodsell, Kaczor-Grzeskowiak, and Dickerson, 1994). The neutron diffraction analysis to 3.0 Å (Arai *et al.*, 2005) has revealed the presence and orientation of water molecules bridging between bases, which contribute to the somewhat complex spine of hydration found in this particular minor groove.

One of the goals of studies on the relationship between DNA sequence and structure has been to relate to DNA sequences in their biological context. This requires knowledge of both native structures and as protein-bound complexes. A number of such native B-DNA crystal structures have been determined that have direct biological relevance, and for them interesting patterns of sequence-dependent flexibility have emerged. A prominent example is that of DNA bending shown by the DNA sequences binding to the human papillomavirus E2 protein (see Sect. 3.6.3). The DNA sequence that binds the transcription factor NF-κB is a 10 base-pair palindromic site, which has been incorporated into the 17-mer sequence, and its crystal structure has been determined (Huang *et al.*, 2005). This shows significant differences compared to the identical DNA sequence when bound to the NF-κB protein (Chen, Ghosh, and Ghosh, 1998). The flanking G:C regions, which have the most contacts with the protein, buckled and opened in the complex but not in the native structure. The large roll angle at the center of the A/T tract in the complex produces significant curvature (see Sect. 3.6.3), but not in the native structure. The narrow minor groove in this structure has a spine of hydration.

A novel host-guest approach has been developed for studying DNA sequences that are unable to crystallize on their own (Coté, Yohannan, and Georgiadis, 2000). This utilizes the N-terminal fragment of the Moloney murine leukaemia virus reverse transcriptase, as a sequence-neutral “host” to bind to a wide range of

16-mer DNA sequences, including ones with mismatches and bound drug molecules (see Sect. 5.6.2). The resulting structures have a dumbbell-like appearance with a protein molecule at each end connected by the oligonucleotide duplex. All crystallize in the same space group, $P2_12_12$, but their sheer size appears to allow for flexibility in the DNA such that sequence-dependent effects are apparent. A study of a DNA analogue of the HIV-1 polypurine A-tract, the asymmetric sequence d(CTTTTAAAGAAAAG) with its complementary sequence, has shown (Coté, Pflomm, and Georgiadis, 2003) that in spite of the presence of the three A-tracts, the structure is essentially straight. Its features are closely similar to several other related sequences in distinct environments, again strongly supportive of the notion that DNA local structure is intrinsic to its sequence rather than to its environment, both as a native structure and often when bound to a sequence-neutral protein. The host–guest approach has been used in a comparative analysis of four self-complementary 16-mer sequences, which has enabled properties of particular dinucleotide steps to be highlighted, such as the positive roll angle of the CpA step (Montaño *et al.*, 2006).

3.3.4 Sequence-Dependent Features of B-DNA: Their Occurrence and Their Prediction

The central question that comparative studies on the large number of DNA crystal structures seek to answer, is whether and to what extent the observed sequence-dependent structural effects are real or are artifacts of (a) crystal-packing effects, with adjacent helices generally having close contacts, or (b) the random and systematic errors that accompany X-ray (and NMR) structures, where the likely errors of medium-resolution structures are in the same range as the differences in some parameters from canonical mean values.

Variations in the local base-pair and step parameters propeller twist, roll, and helical twist for the original Dickerson–Drew structure, the twofold symmetric variant, and several other dodecamer structures are illustrated in Fig. 3.11 (see Chap. 2 for definitions). The sequence-dependent effects are most reproducible for the central 6–8 residues of these sequences since the 3′ and 5′ termini of the dodecamer structures are the regions most involved in packing interactions with other molecules in the crystal lattice. The pattern of base and base-pair morphological parameters is also consistent with those derived from NMR data in solution using the method of residual dipolar couplings (Trantirek *et al.*, 2000). The most significant variations in values for these parameters from canonical values are:

- The local helical twist varies by up to 15° , with pyrimidine-3′,5′-purine steps having lower than average values and purine-3′,5′-pyrimidine steps having higher than average ones. These differences correlate with differences in the susceptibility of the sugar–phosphate backbone bonds in this sequence to be cleaved by the enzyme DNAase I (Lomonossoff, Butler, and Klug, 1981), with the TpC step, which has a high twist, being the most readily cleaved. The mean twist angle in the crystal structures is 36° , in accord with the fiber-diffraction value
- Propeller twists are significantly greater for A•T base pairs than for G•C ones, by an average of $5\text{--}7^\circ$

- Roll angles for pyrimidine-3',5'-purine steps have positive values, since they open up toward the minor groove, whereas purine-3',5'-pyrimidine steps have negative roll angles with major groove opening

There is good evidence that at least some sequence-dependent features seen in the dodecamer structure have direct relevance to "real" DNA sequences, and are not restricted to short crystalline ones, especially the larger-scale features such as the groove-width variation. For example, it has been found (Drew and Travers, 1984) that there are variations in the ability of the DNase I enzyme to cleave the phosphodiester backbone of the 160 base pair *E. Coli tyrT* promotor within runs of A•T base pairs. These variations correlate with changes in minor-groove width that occur in them when TpA steps are present or absent.

The sequence-dependent changes in helical and propeller twist, roll, and slide were initially rationalized on the basis of steric clashes between substituent atoms on individual bases – the Calladine Rules (Calladine, 1982; Calladine and Drew, 1984). These rules are based on the hypothesis that the observed structural effects derive from opposite-strand purine–purine steric hindrance, and thus give an essentially mechanical view of DNA structural features as arising from the interplay of rigid-body Newtonian forces. For purine-3',5'-pyrimidine steps these clashes are between major-groove substituent atoms O6 of guanine and N6 of adenine, and between minor-groove N2 of guanine and N3 of adenine for pyrimidine-3',5'-purine steps. On this basis, little steric clash would be predicted for purine-3',5'-purine and pyrimidine-3',5'-pyrimidine steps. The rules suggest that clashes are avoided by combinations of changes in twist, roll, and slide. Their relative proportions depend on the nature of the bases involved:

- A decrease in local helical twist decreases minor-groove clashes,
- An increase in roll angle in the groove with clashes,
- Separation apart of successive purines by means of an increase in slide between them,
- A decrease in propeller twist. The changes in slide result in alterations in the backbone angle δ , which thus has a higher value for purine compared to pyrimidine nucleosides.

The Calladine Rules, which have reasonable predictive capability for some (though by no means all) B-DNA structures other than d(CGCGAATTCGCG), only have limited applicability to other helical types, especially A-forms where there is inherently high propeller twist. The inability of the rules to take factors such as hydration, electrostatics, and base-stacking into account, does mean that they only provide a first-order approximation to the understanding of sequence-dependent effects, even in B-DNA structures.

The rules provide a means of understanding the structural behavior of dinucleotide base steps, of which there are just 10 distinct types. The rules also imply that each step is largely independent of its neighbors, in reality a considerable oversimplification. Extension to four-base sequences would require information for 136 distinct possible combinations. The establishment of the first B-DNA oligonucleotide structure provided information on only a very small number of these possible nearest-neighbor combinations of DNA sequence, and so there have been concerted attempts to experimentally determine structures for other sequences, both relating to and distinct from the Dickerson–Drew one. Even though several hundred nucleic acid crystal structures

containing B-DNA have been determined, less than half of the 136 tetranucleotide combinations have been sampled experimentally. A theoretical study of sequence-context effects at the tetranucleotide level (Packer, Dauncey, and Hunter, 2000) used an experimental database with 66 out of the 136 included. This study suggests that A/T-containing dinucleotide steps tend to be significantly more context-independent than C/G-containing ones.

There have been a large number of subsequent studies that have examined relationships between parameters and their underlying structural basis, which have been able to use the increasing number of B-DNA structures available in the PDB/NDB by searching for correlations between parameters (Suzuki *et al.*, 1997; Subirana and Faria, 1997). Although the available structures do not enable *all* possible flanking sequences to be examined (see Sect. 3.4), it does appear that they only influence some individual base steps, notably pyrimidine-3',5'-purine ones (Subirana and Faria, 1997). The sequence-dependence of helical twist angles has been examined in a set of 38 structures (Gorin, Zhurkin, and Olson, 1995), which has demonstrated a significant correlation between twist and roll, as a consequence of steric clashes, some of which are in addition to the purine-purine clashes defined by the Calladine Rules (Calladine, 1982). These authors have introduced an empirical clash strength function, which while able to reliably predict the twist of most steps, fails with some CpA ones. An ambitious attempt to crystallographically map out sequence-dependency has been made with the inverted repeat sequence motif d(CCnnnN₆N₇N₈GG), where N₆, N₇, and N₈ are A, T, C, and G (Hays *et al.*, 2005). Out of the 64 permutations 63 were crystallized, resulting in 37 crystal structures from 29 different sequences. This set contained A- and B-DNA structural types, as well as several involving four-stranded Holliday junctions.

Considerations of base-base interactions using empirical energy functions (Hunter, 1993; Hunter and Lu, 1997; Elsayy, Hodgson, and Caves, 2005), have been a fruitful approach to understanding DNA sequence-dependent structure. These energy functions include terms for van der Waals and electrostatic interactions, with charge distributions in bases being calculated (Hunter, 1993; Farwar, Packer, and Hunter, 2006), on the basis of a π -electron model rather than with the more commonly used point-charge approach. Experimental values for rise, roll, and tilt, which are especially dependent on base-stacking, are well reproduced by this approach. An extension of this approach has used a genetic algorithm to find global minima (Packer and Hunter, 2001) for a set of 30 diverse oligonucleotide structures, all of which also have known crystal structures. Most of them are well-predicted, with the sequence-dependent trends in base-step and pair morphologies mostly being well-reproduced. When backbone and base variability is taken into account (Farwar, Packer, and Hunter, 2006), the average rmsd comparing prediction to experiment is 2.24 Å. Since the accuracy of the structures themselves is unknown, correlations with computational predictions of the details of sequence-dependency do need large samples of structures if they are to have statistical reliability.

Empirical energy functions have been used to theoretically examine all 32,896 unique octamer DNA sequences (Gardiner *et al.*, 2003; 2004), and suggest (i) that the overwhelming majority have B-like conformations; A-type duplexes are likely to have consecutive guanosine residues in the sequence and (ii) that structural diversity is much less than sequence diversity. Importantly though, it does appear that there are differences, at least in terms of flexibility, between promoter and nonpromoter sequences.

Backbone conformational angles show considerable variability in B-DNA structures with respect to sequence (Table 3.5). Even within the Dickerson–Drew structure, individual angles typically show a $>45^\circ$ spread of values along the sequence. The central question, relevant both to variations in backbone angles and to base morphological parameters, is whether these differences reflect (i) true sequence-dependent variability, (ii) the influence of crystal-packing interactions, or (iii) are a reflection of a lack of precision in the crystal structure determinations. Surveys of conformational angles (Schneider, Berman, and Neidle, 1997) from a large number of crystal structures show that each angle occupies a discrete range (Table 3.6), with some such as ζ occupying two regions. The angles have Gaussian distributions, suggesting that the variations reflect both experimental inaccuracies and real fluctuations. A subsequent survey (Sims and Kim, 2003) has taken conformational information on 6253 individual dinucleoside phosphates, also from oligonucleotide structural entries in the Nucleic Acid Database. This data has been comprehensively mapped using a multidimensional scaling method, which shows that the data separates into nine major conformational types, corresponding to A-, B-, and Z-type clusters. The distributions of torsion angles correspond well with those found six years previously (Schneider, Berman, and Neidle, 1997), using a smaller data set. Interestingly, analysis of the more recent data set also revealed a number of minor conformational clusters, corresponding to noncanonical features.

The experimental data does not point to clear sequence-dependency relationships between backbone conformation and helical parameters, in broad agreement with a theoretical analysis (Packer and Hunter, 1998), which concludes that roll, tilt, and rise are inherently backbone-independent. Helical twist is dependent on backbone conformation, though not in a straightforward way.

Table 3.5 Ranges of Backbone Torsion Angles ($^\circ$) Observed in High-Resolution (Defined as Better than or Equal to 1.9 Å) B-DNA Crystal Structures (from Schneider, Neidle, and Berman, 1997). The Ranges Corresponding to B_I and B_{II} Phosphate Conformations are Indicated

	α	β	γ	δ	ϵ	ζ	χ
Range	270–330	130–200	20–80	70–180	160–270	150–210 230–300	200–300
Mean values	298	176; B_I 146; B_{II}	48	128; B_I 144; B_{II}	184; B_I 246; B_{II}	265; B_I 174; B_{II}	258; B_I , Pu 241; B_I , Py 271; B_{II} , Py

Table 3.6 Selected A-DNA Crystal Structures

Sequence	PDB number	NDB number
d(GGGCGCCC), tetragonal	2D94	ADH026
d(GGGCGCCC), hexagonal	2D95	ADH027
d(CCCCGGGG)	187D	ADH056
d(ACGTACGT)	243D	ADH070
d(CCGGGCCCGG)	321D	ADJ082
d(GCGTACGTACGC)	117D	ADL046
d(CGCCCCGCGGCG)	399D	ADL047
d(CATGGGCCCCATG)	1DC0	BD0026
d(AGGGGCGGGGCT)	1ZJE	AD0053

3.4 A-DNA Oligonucleotide Crystal Structures

DNA double helices of the A-type are produced in fibers of random-sequence DNA under conditions of low humidity (see Sect. 3.2), and in solution when the water activity is reduced by the addition of various alcohols. The fiber diffraction studies suggest that certain runs of sequence such as alternating G•C base pairs, have a tendency to be in an A form. Single-crystal analyses of particular lengths of oligonucleotide sequences have determined many to be of the A-type, especially when they have high G/C content. However, for the majority of such structures, it is now apparent that this does not reflect so much any inherent structural preference, as more the crystal-packing requirements for these particular lengths of oligonucleotide duplexes. The high-alcohol crystallization conditions commonly used in oligonucleotide crystallography may also be a contributory factor. Nonetheless, these structures have provided invaluable insights into the flexibility and conformational preferences of A-DNA helices in general, even if their significance in biological terms is not always apparent. There is at least one instance where A-type structures are of undoubted importance: this is in the case of RNA–DNA hybrids, formed, for example, during transcription and replication. Since RNA helices are always constrained to be in an A form, then DNA–RNA hybrids would be expected to be similarly constrained. This has been found to be the case by both fiber-diffraction studies of RNA–DNA hybrid polynucleotides and by single-crystal analyses. For example, the structure of the duplex formed by r(GCG)d(TATACGC) shows that the DNA strand has adopted a conformation close to that of duplex RNA, with an 11-fold helix (average helical twist of 33°), C3'-*endo* sugar pucker and an A-type helical backbone conformation (Wang *et al.*, 1982).

3.4.1 A-Form Octanucleotides

A large number of self-complementary octanucleotides, of widely varying sequence type, have been crystallized and found to have A-DNA family structures. As with B-type oligomers, structural features averaged over a number of such sequences are close to those of fiber-diffraction A-DNA. By contrast with B-DNA oligomer crystal structures, backbone torsion angles tend to have narrow ranges of values, reflecting the greater conformational rigidity of the A-form. Some of the A-form octamer sequences crystallize in the tetragonal space group $P4_32_12$; most of the others are in the hexagonal one $P6_1$. It has been possible to vary conditions such that some sequences such as d(GGGCGCCC) (Shakked *et al.*, 1989) and d(GTGTACAC) (Jain, Zon, and Sundaralingam, 1991) have been crystallized in both forms. The hexagonal form has also been found for sequences such as d(GGGGCCCC) (McCall, Brown, and Kennard, 1985), d(GGGTACCC) (Eisenstein *et al.*, 1990), and d(GGGATCCC) (Lauble *et al.*, 1988) as well as for a number of mismatch variants. All have helical and conformational features within the general A class, but with some marked local variations, in particular in the overall helical repeat, in minor-groove widths and in base-step parameters. The wide minor groove, of ~15 Å in d(GGGGCCCC) contrasts with the more typical value of 9.7–9.8 Å in d(GGGTACCC), which is close to that for canonical fiber-diffraction A-DNA, of 11.1 Å. Major-groove width cannot be fully assessed in an octanucleotide duplex, since eight rather than seven phosphate groups are required in order to calculate the shortest phosphate–phosphate interstrand distance.

Approximated major-groove widths show very wide variation, ranging from 5 Å in the low-temperature form of d(GGGCGCCC) (Clark *et al.*, 1990) to over 12 Å in d(GGGGCCCC) (McCall *et al.*, 1986), compared to the very narrow value of 2.2 Å from fiber diffraction. The tetragonal form, with sequences having a pyrimidine-3',5'-purine step at positions 4 and 5, that is at the center of the helix, shows significant deviations from standard A-form backbone geometry at this point, with torsion angles α and γ having *transoid* values rather than the normal *gauche* and *gauche* ones respectively. This discontinuity has been observed in several structures, for example that of d(ATGCGCAT) (Clark *et al.*, 1990). It is likely that this effect is a consequence of the requirements imposed by hydration and crystal packing in the tetragonal unit cell, rather than being an intrinsic property of these sequences.

The 9-base-pair binding site of the Sp1 transcription factor is GC-rich, and three crystal structures of this sequence, d(GGGGCGGGG) embedded within 12 or 13 base-pair sequences have been reported, with all having the ends participating in overhanging base pairs in the crystal lattice (Dohm *et al.*, 2005), and all forming A-type duplexes. There is some variation in detailed conformation between them, even though all have such standard A-type features as 11 base pairs per helical turn, again indicative of a degree of local flexibility that is markedly less than for B-type structures.

3.4.2 Do A-Form Oligonucleotides Occur in Solution? Crystal-Packing Effects

Are crystal-packing factors the driving force behind *all* octanucleotide structures being in the A form? Solution NMR studies on the sequence d(ATGCGCAT) (Clark *et al.*, 1990) have demonstrated that it shows B-family behavior in solution even though the crystal structure is unequivocally of the A-type. This dichotomy between crystal and solution environmental constraints is also shown by the sequence d(GGATGGGAG), which forms part of the binding site for the transcription factor TFIIA that forms part of the initiation complex for transcription of the 5S RNA gene in *Xenopus*. The (low-resolution) 3 Å crystal structure of the duplex formed by this sequence and its complementary strand, shows an A-family helix (McCall *et al.*, 1986), with average major-and minor-groove widths of 14.8 and 15 Å respectively, and a helical repeat of 11.5 base pairs. These values closely correspond to those for A'-RNA from fiber diffraction studies. Nuclease digestion studies, showing a repeat of 11.4 base pairs per turn, together with circular dichroism measurements (Fairall, Martin, and Rhodes, 1989), are consistent with the assignment of a structure for the d(GGATGGGAG) sequence as well as for the complete 54 base-pair TFIIFA binding site, that is not classical B-DNA. On the other hand, NMR and other biophysical methods all indicate that the nonamer in solution has normal C2'-*endo* sugar puckers and B-DNA range glycosidic angles (Aboul-ela *et al.*, 1988).

Short A-DNA double helices are insufficient in length for the major groove to be fully formed, and so estimates of its dimensions are necessarily approximate. X-ray analyses of two dodecamers (in two distinct space groups) have each revealed the structure of a complete turn of A-DNA double helix, as well as conformational features that are not subject to the crystal-packing factors of the octanucleotides. The structure of the sequence d(CCCCCGCGGGG), with both alternating and non-alternating nucleotides (Verdaguer *et al.*, 1991) has average backbone conformational angles and helical parameters that are remarkably close to those in A-DNA fibers,



Figure 3.14 Structure of a complete turn of A-form DNA, as found in d(CCCCCGCGGGG) (Verdaguer *et al.*, 1991).

with some local backbone angle variations, for example *trans* α and γ angles at one of the two CpG steps (Fig. 3.14). This structure, together with others such as that of the A-DNA decamer d(ACCGGCCGGT) (Frederick *et al.*, 1989), is consistent with the notion that very GC-rich sequences have a higher tendency to form A-type duplexes. The sequence d(CCGTACGTACGG), of which two-thirds consists of G•C base pairs, likewise has an A-DNA structure (Bingman, Zon, and Sundaralingam, 1992), this time without any *transoid* α/γ angles. The average major-groove width in this structure, of 3.5 Å, is similar to the 2.2 Å value from fiber diffraction. A high-resolution crystallographic study of the sequence d(AGGGGCCCT) in two distinct space groups concludes (Gao, Robinson, and Wang, 1999) that conformation is less conserved in this circumstance than when different A-type sequences crystallize in the same space group.

Changes in crystallization conditions do not necessarily result in changes to the resulting helical structure. This has been demonstrated in a study (Finley and Luo, 1998) of the structure of the sequence d(GACCGCGGTC), crystallized under two contrasting conditions of ionic strength. This sequence, which is half of the human papilloma virus E2 binding site, forms an A-type helix with closely similar helical parameters under both conditions. However there is a marked difference in the degree of helical bending, with that grown under higher-salt conditions being bent by 31° in the major-groove direction compared to 15° for the low-salt form. Such bending may be relevant to the need for flexibility when particular sequences bind to their cognate proteins, as we shall see in Chap. 7. Hydration in A-DNA oligonucleotides is generally extensive, with observations of phosphates and bases hydrogen-bonding to water molecules (Tippen and Sundaralingam, 1997). Most solvent ordering occurs in the major groove, with polygons of water molecules being observed in some crystal structures (Eisenstein and Shakked, 1995). These polygon arrangements are especially apparent at high resolution (Egli *et al.*, 1998).

3.4.3 The A $\leftrightarrow$ B Transition in Crystals

There are now over 150 A-DNA crystal structures in the NDB, which provide a very detailed view of this DNA form and the myriad of deviations from canonical ideality (notably in major-groove width) that are possible (Wahl and Sundaralingam, 1997). It may be thought that there is then little more to learn about A-DNA. However the capacity of DNA structure to surprise is still apparent. The crystal structure of the dodecamer d(CGCCCGCGGGGCG) is remarkable in that parts of the structure demonstrate true hybrid character (Malinina *et al.*, 1999). The central octamer region has an A-DNA conformation, but with a high degree of bending around the central purine-3',5'-pyrimidine base step (Fig. 3.15). This 65° bending results in very close phosphate groups. By contrast the terminal nucleotides have many B-DNA characteristics, such as B-like sugar puckers, backbone conformational angles, and base slide. This structure thus has features akin to an A $\leftrightarrow$ B transition, albeit with distinct conformations at different points along the sequence. Whether this structure is a consequence of the influence of crystal-packing forces or innate tendencies of this sequence is not clear.

The high-resolution structure of d(CATGGGCCCATG) has features that are throughout its length intermediate between A and B (Ng, Kopka, and Dickerson, 2000), and thus resembles a trapped intermediate in the A $\leftrightarrow$ B transition. There is a significant base shift, producing an A-like channel in the center of the helix, and groove widths are intermediate between the A and B ideal values. A complementary view is provided by the analysis of six d(GGCGCC) duplex hexanucleotide crystal structures with varying degrees of cytosine bromination or methylation (Vargason, Henderson, and Ho, 2001; Ng and Dickerson, 2001). This results in differing crystal forms and structures that progressively show features that span A, B, and plausible intermediate states. For example, the major groove becomes progressively deeper and values for the slide parameter become increasingly negative on proceeding from the B-type to the A-type structures. It has been suggested (Ng and Dickerson, 2002) that a factor in the A/B continuum seen in the d(CATGGGCCCATG)



Figure 3.15 Structure of the A/B hybrid d(CGCCCGCGGGGCG), with the large bend at the center of the structure being apparent (Malinina *et al.*, 1999).

crystal structure is the tendency of GpG steps in this sequence to prefer A-like features.

3.5 Z-DNA – Left-Handed DNA

3.5.1 The Z-DNA Hexanucleotide Crystal Structure

The first oligonucleotide duplex single-crystal structure to be solved (by multiple isomorphous replacement methods) was that of the alternating pyrimidine–purine sequence d(CGCGCG) in 1979, at the very high resolution of 0.9 Å (Wang *et al.*, 1979). This structure is remarkable in that even though it consists of a duplex formed by the two antiparallel hexamer strands, the helix is (quite unexpectedly) a left-handed one (Fig. 3.16). The backbone is irregular compared to A or B-DNA since the dC and dG residues have very distinct conformations (Fig. 3.6), resulting in a “zig-zag” arrangement of phosphate groups – hence the helix has been termed Z-DNA. The same left-handed arrangement has also been found (Arnott *et al.*, 1980) in fibers of the alternating polynucleotide poly(dC-dG)•poly(dC-dG) – see Sect. 3.2.2. Z-DNA was in effect discovered some years earlier during the course of circular dichroism studies on poly (dG-dC)•poly(dG-dC), when it was observed (Pohl and Jovin, 1972) that on increasing the ionic (as salt) concentration beyond ~4M NaCl, the CD spectrum became inverted from its standard B-DNA-associated shape. This major spectral change indicates a transition

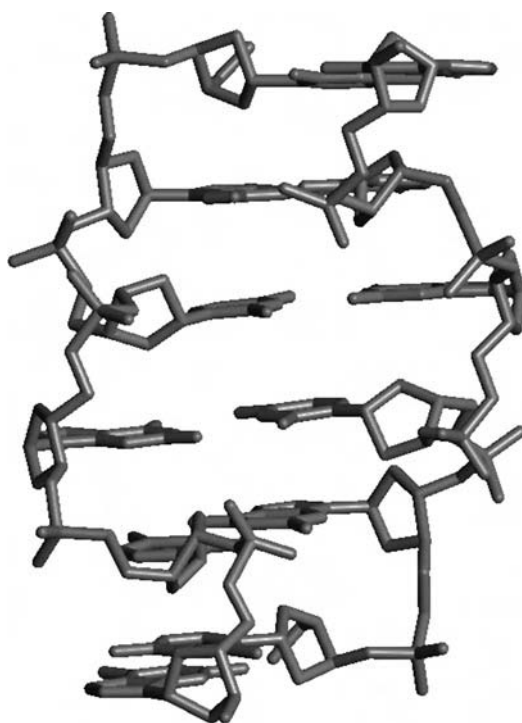


Figure 3.16 Structure of the Z-DNA sequence d(CGCGCG) (Wang *et al.*, 1979).

to a quite different conformational form, which is now accepted to be Z-DNA. A number of other physico-chemical techniques have subsequently been used to study this left-handed structure in addition to X-ray crystallography (Rich, Nordheim, and Wang, 1984).

3.5.2 Overall Structural Features

Z-DNA oligonucleotides (and the poly(dG-dC)•poly(dG-dC) polynucleotide) have the following conformational characteristics (Table 3.7):

- The purine (deoxyguanosine) nucleosides have *syn* glycosidic angles with a χ range of 55–80° (mean 60°), together with C3'-*endo* sugar puckers. The conformation about the α/γ backbone torsion angles is *gauche*⁺, *trans*, and *gauche*⁻ about angle ζ .
- The pyrimidine (deoxycytidine) nucleosides have *anti* glycosidic angles with a χ range of -145 to -160° (mean -152°), and C2'-*endo* sugar puckers. The α/γ conformation is *trans*, *gauche*⁺ and that about angle ζ is *gauche*⁺.
- The G•C base pairs are of standard Watson-Crick type.

A consequence of these differences between purine and pyrimidine nucleosides is that the helical repeating unit is forced to be the CpG dinucleotide, rather than the mononucleotide one in standard right-handed canonical A- and B-DNA. The distinct α/ζ phosphate conformations of the two nucleosides result in the characteristic zigzag appearance of the backbone. Detailed examination of d(CGCGCG) in various crystal forms (Wang *et al.*, 1981) has shown that there is a secondary backbone conformational family with a distinct set of phosphate orientations. This secondary conformation is termed Z_{II}, with the standard type as described above being termed Z_I. The Z_{II} conformation results in purines having a *gauche*⁺ rather than a *gauche*⁻ value for the ζ backbone torsion angle and a *gauche*⁻ rather than a *trans* value for angle ϵ . Z_{II} pyrimidines have (rather smaller) changes in angles α and β compared to Z_I ones. Occurrence of the Z_{II} conformation at a particular residue is probably related to the coordination of the phosphate group at this point to a hydrated magnesium ion (Wang *et al.*, 1981; Egli *et al.*, 1991) – Z-DNA oligonucleotides are usually crystallized in the presence of magnesium and spermine ions. Thus, the Z_{II} conformation does not occur at this point in the sequence in the pure-spermine, magnesium-free structure of d(CGCGCG). Neutron diffraction studies on d(CGCGCG) have shown that the water molecules in the Z-DNA minor groove are well-ordered whereas those in the major groove are much less so (Chatake *et al.*, 2005).

Table 3.7 Selected Z-DNA Crystal Structures. The 2nd Strand is not Listed for Sequences that are not Self-Complementary

Sequence	PDB number	NDB number
d(CGCGCG) at ultra-high resolution	110T	ZD0004
d(CGCGCG) + Mg ²⁺ , spermine	2DCG	ZDF001
d(GCGCGCGCGC)	279D	ZDJ050
d(CACGTG) + propanediamine	2F8W	ZD0016
d(GTCGCGCGCCATAAACC)	2ACJ	PD0693

3.5.3 The Z-DNA Helix

The Z-DNA double helix is to a large degree represented by the structure of the hexanucleotide duplex d(CGCGCG); however there are slight differences between individual CpG units, quite apart from the occurrence of Z_I and Z_{II} forms. Idealized helices for both of these have been generated (Wang *et al.*, 1981). The fiber diffraction model for Z-DNA is closest to the Z_I helix. This Z helix has a 44.6 Å pitch with 12 base pairs per helical turn and a diameter of ~18 Å (Fig. 3.7), making it somewhat slimmer than a B-DNA helix (with a diameter of ~20 Å). The helical twists for two base-pair steps CpG and GpC are quite different as a consequence of the asymmetry in guanosine and cytidine conformations, with the CpG step having an exceptionally small twist and almost no base-stacking between the two C•G base pairs in the step. A further consequence of the *syn* guanosine conformation is that the base pairs are not positioned astride the helix axis, as in B-DNA. Rather, their edges are at the surface of the helix. The N7 and C8 atoms of the imidazole 5-membered ring of guanine and (to a lesser extent) the C5 atom of cytosine, actually protrude onto the helical surface of what would be the major groove, making the surface slightly convex at this point (Fig. 3.17). In other words, Z-DNA does not have a major groove at all. Its minor groove is very narrow and deep, and lined with phosphate groups (Table 3.3). The differences in phosphate orientation between Z_I and Z_{II} helices result in the latter having a ~1 Å wider minor groove.

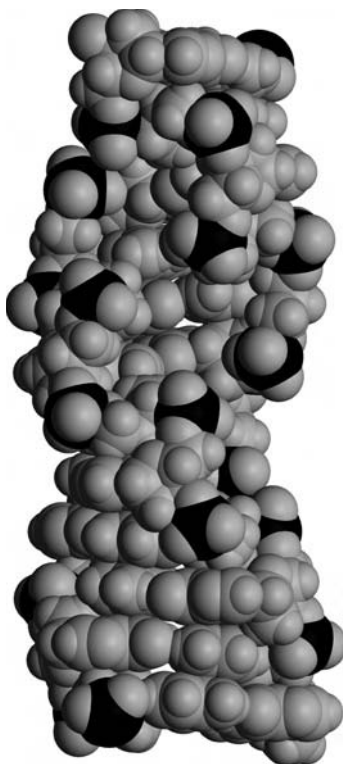


Figure 3.17 Van der Waals surface representation of a Z-DNA left-handed duplex, with phosphorus atoms highlighted in black.

3.5.4 Other Z-DNA Structures

Z-DNA oligonucleotides have been crystallized with a range of related sequences, including pure analogues such as d(CGCGCGCG) (Fujii *et al.*, 1985) and d(CGCGCGCGCG) (Ban, Ramakrishnan, and Sundaralingam, 1996), as well as variants where changes have been made to base and sequence type. All show the resilience of the Z-DNA structural entity. The central CpG can be replaced by TpA whilst maintaining the left-handed structure (Wang *et al.*, 1984). This shows that Z-DNA can tolerate some A•T base pairs, although they do tend to destabilize the structure. The cytosines are required to be 5-methylated in order for stabilization of the d(CGATCG) sequence, and in general modified cytosines are necessary if a Z-DNA structure contains A•T base pairs. It has been suggested (Wang *et al.*, 1984) that A•T base pairs are not able to take part in the ordered Z-DNA groove hydration, which plays an important role in maintaining the integrity of the Z-DNA structure (Egli *et al.*, 1991), by contrast with G•C base pairs. Hence, A•T base pairs cannot by themselves form a Z-DNA structure. The slight preference of guanosine nucleosides (unlike adenosine) to adopt the *syn* rather than the *anti* glycosidic conformation (see Chap. 2) is also a factor in the GC preference of Z-DNA. It is then surprising that reversal of the central pyrimidine-3',5'-purine sequence (i.e. so that it is no longer purely alternating) in the structure of d(CGATCG), still retains a Z-DNA type structure (Wang *et al.*, 1985), albeit with the thymidines and adenines adopting *syn* and *anti* glycosidic conformations respectively. This energetically unfavorable Z structure was only stabilized by having the cytosine bases methylated or brominated at the 5 position of the base. Nonetheless, its existence suggests that the requirement for the formation of a Z-DNA structure is not so much that the sequence should be an alternating C/G-rich one, but that it retains the key feature of alternating *syn* and *anti* nucleosides. It is likely that replacement of thymine by uracil would increase Z-form stability, since the methyl group of the former presents a significance hydrophobic hindrance to solvent-ordering around the Z helix. This has been borne out by subsequent structural studies on uracil-containing Z sequences (Schneider *et al.*, 1992). In general nonalternating sequences can form Z-DNA structures in the crystalline state when the nonalternating pyrimidines are able to adopt a *syn* conformation (Basham, Eichman, and Ho, 1999).

The structure of the B-Z junction has been revealed in a crystal structure (Ha *et al.*, 2005) of a 15 base-pair sequence containing within it a CG hexanucleotide Z-DNA motif. This was tethered in the Z-form by being bound to domains of a Z-DNA-binding protein, with the effect that eight base pairs are in the Z-form and six are in a B-DNA conformation (Fig. 3.18). Strikingly, one A•T base pair at the junction is not formed and the adenine and thymine bases are displaced so that they are flipped out into the exterior of the helix, presumably in order to relieve the steric strain at the junction. This has enabled the sharp B-Z backbone turn at the junction to be formed, and continuous base-stacking to be preserved throughout the length of the duplex. It is notable that the resulting B-helix is of a standard regular conformation.

3.5.5 Biological Aspects of Z-DNA

The accessibility of the C8 (and to a lesser extent the N7) position of guanine in Z-DNA renders it susceptible to attack by several electrophilic compounds that have a



Figure 3.18 Structure of an oligonucleotide having a B-Z junction (Ha *et al.*, 2005).

preference for these positions. Examples include the N-aryl carcinogens acetylaminofluorene and aflatoxin, as well as C8-bromination. All stabilize Z-DNA and promote the B–Z structural transition. Thus, Z-DNA tracts within genomic sequences could act as mutational hot spots for these agents. The fact that Z-DNA formation is also facilitated by methylation at the C5 position of cytosine, a possible mechanism in some eukaryotic organisms for regulating gene transcription, suggests that Z-DNA could play a role in gene regulation.

Z-DNA is a higher-energy polymorph of DNA and in its linear form is less stable than either A- or B-forms, requiring high-salt or high-alcohol concentrations

for maintenance of its structure in solution. These reduce electrostatic interactions between interstrand phosphate groups, which in Z-DNA are much closer together than in B-DNA (7.7 Å in the Z₁ form compared to 11.7 Å across the minor groove in B-DNA). Z-DNA sequences can also be stabilized at physiological salt levels when incorporated in underwound, negatively supercoiled covalently closed-circular DNA, for example in plasmid DNA. Even low (~1–2%) levels of Z-forming sequence can significantly change supercoiling properties. It has been proposed (Rahmouni and Wells, 1992) that a role for Z-DNA *in vivo* is to act as a signal for the induction of transcription via supercoiling.

The classic Z-forming sequence d(CG)_n is much rarer in eukaryotic than in prokaryotic organisms. However, the sequence d(CA/GT)_n, which can also form left-handed DNA in negatively supercoiled plasmids, is much more prevalent in eukaryotic genomes. The overriding questions of whether Z-DNA actually exists naturally, and if so what its function may be, has not as yet been fully answered (Rich and Zhang, 2003). It is on balance, likely though that Z-DNA can be formed in cells under appropriate circumstances. A number of proteins that bind to Z-DNA have been isolated, and several cocrystal structures have been determined. Following the crystal structure analysis of a Z-DNA complex with the Zα high-affinity binding domain of an RNA-editing enzyme (Schwartz *et al.*, 1999), a subsequent database search for other proteins with the same Z-binding motif found a closely similar domain in a tumor-associated protein (DLM-1). The crystal structure (Schwartz *et al.*, 2001) of a Z-DNA complex with this shows closed structural analogy with the Zα complex. In both the critical protein to Z-DNA contacts are made by several conserved amino acid residues. The crystal structure of a Z-DNA complex with a pox virus protein that also contains the Z-binding motif shows nine amino-acid residues in contact with the DNA (Ha *et al.*, 2004). In all these instances the nature of the protein suggests that Z-DNA plays a role in regulation of transcription/translation.

3.6 Bent DNA

3.6.1 DNA Periodicity in Solution

DNA under physiological solution conditions is generally assumed to be in a B-type conformation, even when combined with chromosomal proteins. There are at first sight some important differences between the B structures as found in the crystal and fiber, and that in solution, even as naked DNA. The structural studies all point to a helical repeat of 10.0–10.1 residues per turn for averaged-sequence DNA. However, plasmid mobility, nuclease digestion, and hydroxyl radical cleavage experiments (Tullius and Dombroski, 1985) have given a rather different value, of close to 10.5 base pairs per turn, again averaged over many base pairs. This apparent discrepancy is only partly resolved by the observation that helical periodicity in solution is markedly sequence-dependent (Rhodes and Klug, 1981). Cleavage at a particular point along a DNA sequence depends on the local helical twist and hence groove width, in a manner that correlates with the features that have been found in the Dickerson–Drew dodecamer, both in the crystal and in solution by NMR.

3.6.2 A-Tracts and Bending

The DNA double helix is usually thought of as a straight molecule. The original Watson-Crick model is of a straight helix, since it was based on data from “straight” DNA fibers. In fact, it must be bent or curved in many of its functional roles, especially when interacting with particular proteins. DNA is even deformed from linearity in many of the dodecamer crystal structures, being typically bent by about 19°. There has been much speculation on the nature of DNA bending in such situations as in the nucleosome component of chromatin (which has 145 base pairs wound around a histone protein core) and closed-circular DNA, prior to determinations of nucleosome crystal structures. Some sequences have an inherent tendency to bend independently of there being any protein present. The best-studied such sequences are those comprising runs of adenines on one strand and thymines on the other (variously termed A-tracts or A/T tracts).

Intrinsic DNA bending and curvature was discovered (Marini *et al.*, 1982; Crothers and Shakked, 1999) in restriction fragments of kinetoplast DNA, which have extensive repeats of A-tracts within them, each having about 5–7 base pairs. These fragments showed highly anomalous gel electrophoresis behavior, with very slow migration through the gel, as if the DNA had a much higher molecular weight than was the case. Bending occurs when sequences or motifs such as 5'-AAAAAA are repeated in phase with the DNA helical repeat itself (Crothers, Haran, and Nadeau, 1990; Hagerman, 1990), that is every 10 or 11 base pairs. The phasing is critical since it forces the tract to be always on the same side of the helix, thereby enabling the individual small bends associated with each tract to add up to a significant amount of total bending. It has been estimated (Koo *et al.*, 1990) that an individual (A)₆ tract bends a helix by about 17–21°. A typical bending sequence, from kinetoplast DNA, is:

5'-.....CCAAAAATGTCAAAAAATAGGCCAAAAATGCCAAAAAT.....

Two distinct structural models have been suggested to account for these observations of bending. In one, a combination of gel retardation and NMR studies have been used (Nadeau and Crothers, 1989), resulting in a structural model that has bending toward the minor-groove direction of the helix. The compression and closure of the minor groove is in part a result of negative A•T base-pair tilt within the A-tract itself. In this, the “junction” model, bending is caused by the abrupt change in structure when the A-tract meets the more normal (straight) B-DNA intervening sequence, resulting in the helix axes of the two segments being at an angle to each other (Fig. 3.19a). Abrupt changes in either tilt or roll can produce this situation. Thus, in the junction model, the A-tracts are themselves not bent. By contrast, the “wedge” model (94) has continuous bending along the A-tract (Fig. 3.19b) as a result of changes in roll at each dinucleotide step – this model has also been suggested for the continuous (induced) bending of general-sequence DNA when wrapped around histones in nucleosomes. The structural basis for bending has been controversial, with there being continuing debate on the roles played by base-pair propeller twist, roll, buckle, and slide both within and flanking A-tracts themselves.

The advent of relevant crystal and NMR structures is providing structural data for distinguishing between these models, with increasing reliability for the various base and base-pair morphological parameters, although no bent structure is as yet available that even approaches atomic resolution.

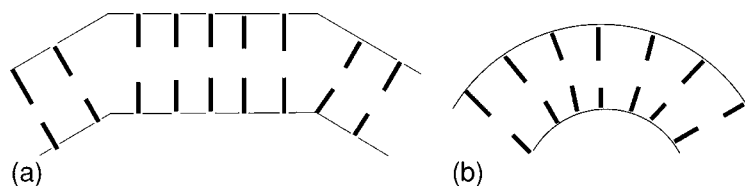


Figure 3.19 Schematic representations of bending models for A tract DNA. (a) junction model; (b) wedge model.

3.6.3 Structures Showing Bending

Several oligonucleotide crystal structures contain short dA•dT sequences that might be thought of as models for A-tracts, as discussed in Sects. 3.1 and 3.2. These have narrow minor grooves with ordered water molecules, high propeller twists for the A•T base pairs, together with, in two cases (Nelson *et al.*, 1987; DiGabriele and Steitz, 1993) bifurcated three-center hydrogen bonds involving adjacent A•T base pairs (Fig. 3.12). Such hydrogen-bonding was however, not observed (Edwards *et al.*, 1992; Woods *et al.*, 2004) in the higher resolution structures determined for d(CGCAAATTTGCG). The (A)₆•(T)₆ tract in one crystal structure (Nelson *et al.*, 1987) is itself a little bent, lending support to the junction model. All three crystal structures have an overall bending toward the major-groove direction, in contrast to the inference from gel studies, that A-tracts are actually bent toward the minor groove. Crystal-packing forces have been invoked to account for this discrepancy. More recently determined structures (see below) are probably better models for A-tracts.

There is some evidence that A-tracts do not inherently require three-center A•T hydrogen bonds at all positions for curvature to be produced, in addition to the structures mentioned above. For example, sequences having inosine•cytosine (I•C) base pairs (these cannot form major-groove three-center hydrogen bonds since inosine lacks a N6 group), still result in bent structures (Diekmann *et al.*, 1992). A combined crystallographic and gel electrophoresis study has been undertaken on a number of A-tract model oligonucleotides with varying numbers of I•C base pairs in the tract (Shatzky-Schwartz *et al.*, 1997). This showed that high propeller twist and bending is maintained when there is isolated I•C substitution for A•T, but abolished for pure I-tracts. Interestingly, this analysis also suggested the existence of attractive three-center NH–H interactions (Luisi *et al.*, 1998).

The degree of bending in a DNA structure can be calculated with several of the helix morphology programs outlined in Chap. 2. The use of normal vector plots is especially useful for visualizing bending (Goodsell and Dickerson, 1994; Dickerson, 1998). This procedure uses the unit vectors perpendicular to the least-squares plane from a base pair, and enables both local and global bending to be analyzed. The method was used to show that the Dickerson–Drew dodecamer (Wing *et al.*, 1980; Drew and Dickerson, 1981; Dickerson and Drew, 1981) is bent by 19°, albeit in the major-groove direction. The crystal structure of the decamer d(CATGGCCATG) shows distinct curvature features. The helix is not straight but has a smooth 23° bend over the four central base pairs (Goodsell *et al.*, 1993). This intrinsic bend is consistent with a straight A-tract model, although the roll compression of the major groove suggests a third type of model distinct from the wedge or junction models. The crystal

structure (Hizver *et al.*, 2001) of the dodecamer d(ACCGAATTTCGGT) is nonisomorphous with the Dickerson–Drew dodecamer, since it crystallises in the triclinic space group P1 rather than the normal $P2_12_12_1$ one for dodecamers. This sequence is a biologically relevant one – it is the high-affinity A-tract binding site of the E2 papillomavirus protein. There are three independent molecules in the crystallographic asymmetric unit, providing three independent views of the preferred structure of this sequence. All three are closely similar, with root-mean-square deviations ranging from 0.5 Å to 0.7 Å, strongly suggesting that their structure represents a global minimum. They have an average curvature of 9°, with very narrow minor grooves in the 5'-AATT regions and highly twisted A•T base pairs. The magnitude and direction of curvature is in accord with solution data on this sequence. The bending is produced by a combination of minor-groove compression in the A-tract and major-groove compression of the flanking C•G base pairs. These are a consequence of the inherent features of the central octamer duplex sequence – in particular high propeller twists for the A•T base pairs and buckling at the A-tract junctions.

The structure of the duplex formed by the non-self-complementary dodecamer d(GGCAAAAACGG), bound to its complementary strand d(CCGTTTTTTGCC) has been analyzed by NMR methods (MacDonald *et al.*, 2001). This A-tract structure is bent by 19°, in accord with gel retardation measurements. The A•T base pairs in the A-tract have high propeller twist, with most of the bending occurring in the short flanking C/G sequences. The A-tract itself in this structure has a small amount (5°) of intrinsic bend. The junction bends have been ascribed to combinations of tilt and roll. NMR methods have also shown (Hud, Sklenar, and Feigon, 1999) that ammonium ions are bound in the A-tract minor groove, and may play a role in the bending. These ions cannot be distinguished from water molecules in oligonucleotide structures by current X-ray methods. The NMR structure of the same decamer duplex as that reported by MacDonald *et al.*, (2001) has also been analyzed by NMR methods (Barbič, Zimmer, and Crothers, 2003). This study has found an average global bend of ~9° into the minor groove for the inner eight base pairs (Fig. 3.20), with roll and tilt from several base pairs contributing to this value. An NMR-based comparison has been made between A-tract sequences with a central ApT and TpA step (Steffl *et al.*, 2004). This has concluded that whereas the large negative roll at the ApT step produces a bend toward the minor groove that are in phase with the bends at the A-tract junctions, the TpA step has a positive roll with bending toward the major groove, cancelling out the junction bends so that there is no overall bend for this sequence (Table 3.8).

It is striking that the macroscopic degree of bending observed in these structures accords reasonably well with solution measurements and with the overall features of the junction model. Some features are common to all the structures such as a very narrow minor groove in the A-tract coupled with high A•T base-pair propeller twists. Yet the fine detail of how bending is achieved at and close to the junctions, as well as the precise degree of bending, differ from one structure to another (Dickerson, 1998). This suggests that there is inherent flexibility of the junctions to A-tract structures, and that there are several equi-energetic bent arrangements that can be readily accessed, albeit in differing ways. This is supported by a survey of bending both in native oligomers and in DNA-protein complexes (Young *et al.*, 1995), which also finds evidence for the straight A-tract model. There has been some debate as to whether, and to what extent, the bending in the crystallographically derived structures is affected by the near-universal use of 2-methyl-2,4-pentanediol (MPD) in crystallization experiments,



Figure 3.20 One of the ensemble of NMR-derived structures (Barbič, Zimmer, and Crothers, 2003) for the duplex formed by d(GGCAAAAACGG) together with its complementary strand d(CCGTTTTTTGCC). This structure has an average global bend of $\sim 9^\circ$ into the minor groove for the inner eight base pairs.

Table 3.8 Selected Bent DNA Crystal and NMR Structures. The 2nd Strand Sequences are not Listed for Non-Self-Complementary Sequences

Sequence	PDB number	NDB number
d(CGCAAAAAGCG)	1D98	BDL006
d(CGCAAAATGCG)	1BDN	BDL015
d(GCAAAATTTTGC)	1RVH	-
d(CGTTTTTAAACG)	1RVI	-
d(GGCAAAACGG)	1NEV	-

where its role is to act as a precipitating agent by decreasing the aqueous solubility of oligonucleotides. Electrophoresis mobility (Sproun *et al.*, 1995) and free-radical cleavage (Ganunis, Guo, and Tullius, 1996) experiments in the presence of MPD suggest that high concentrations do produce a change in A-tract structure and decreased bending, although these “low resolution” methods clearly cannot specify the molecular nature of the changes. On the other hand, the majority of the A-tract model crystal structures do actually show solution-relevant bending, even though MPD may in some circumstances qualitatively affect its degree (Dickerson, Goodsell, and Kopka, 1996). Recent simulation studies (Rohs, Sklenar, and Shakked, 2005) using a Monte Carlo simulation approach on oligonucleotide sequences that bend when bind to the E2 protein from papillomaviruses, have resulted in excellent correspondence with the earlier crystal structures (Hizver *et al.*, 2001).

3.6.4 The Structure of Poly dA•dT

The structure of poly dA•dT itself has been the subject of a number of studies. Fiber diffraction methods have defined a structure, the “heteronomous” model (Arnott

et al., 1983) having distinct backbone conformations and sugar puckers for the two dA and dT strands. The results of subsequent analyses (Aymami *et al.*, 1989; Chandrasakaran and Radha, 1992), although differing in detail, are consistent with this structure, with both strands having B-like conformations, albeit with a narrow minor groove. Propeller twists for the A•T base pairs are likely to be significant ($\sim 15^\circ$), but not as high as in the earlier crystal structures, thus reducing the possibility of three-center hydrogen-bonding. Molecular dynamics studies (Fritsch and Westhof, 1991; Sherer *et al.*, 1999; Sprous, Young and Beveridge, 1999; Strahs and Schlick, 2000) similarly indicate that this type of hydrogen-bonding is not of primary importance in stabilizing oligo or poly dA•dT, but merely that it can occur when the geometric circumstances allow.

3.7 Concluding Remarks

Fiber-diffraction studies on DNA polynucleotides have long established the inherent flexibility of the double helix. These have also defined the major polymorphs in terms of A, B and Z duplexes. The fact that the majority of polynucleotide fiber structures can be readily interconvertible under appropriate environmental conditions has also shown us that they actually represent low-energy points along a continuum of structures. The very large number of oligonucleotide crystal structures now known are mostly representative of this continuum, with the important caveat that it is sometimes not possible to distinguish between true fluctuations from the canonical mean, and experimental uncertainties, especially when these are not fully defined. The aim of establishing a straightforward code for correspondence between sequence and structure remains elusive, and indeed may not exist. Instead, these structures are telling us much about DNA flexibility, its low-energy states and the pathways between them. Succeeding chapters describe how this flexibility can be exploited by interactions with other molecules, small and large.

References

- Aboul-ela, F., Varani, G., Walker, G. T., and Tinoco, I., Jr. (1988). The TFIIIA recognition fragment d(GGATGGGAG).d(CTCCCATCC) is B-form in solution. *Nucleic Acids Res.*, **16**, 3559–3572.
- Abrescia, N. G. A., Thompson, A., Huynh-Dinh, T., and Subirana, J. A. (2002). Crystal structure of an antiparallel DNA fragment with Hoogsteen base pairing. *Proc. Natl. Acad. Sci. USA*, **99**, 2806–2811.
- Arai, S., Chatake, T., Minezaki, Y., and Niimura, N. (2002). Crystallisation of a large single crystal of a B-DNA decamer for a neutron diffraction experiment by the phase-diagram method. *Acta Crystallogr.*, **D58**, 151–153.
- Arai, S., Chatake, T., Ohhara, T., Kurihara, K., Tanaka, I., Suzuki, N., Fujimoto, Z., Mizuno, H., and Niimura, N. (2005). Complicated water orientations in the minor groove of the B-DNA decamer d(CCATTAAATGG)₂ observed by neutron diffraction experiments. *Nucleic Acids Res.*, **33**, 3017–3024.
- Arnott, S. (1999). In *Oxford Handbook of Nucleic Acid Structure* (Ed. Neidle, S.), p. 1, Oxford University Press, Oxford.
- Arnott, S., Chandrasekaran, R., Birdsall, D. L., Leslie, A. G. W., and Ratliffe, R. L. (1980). Left-handed DNA helices. *Nature*, **283**, 743–745.
- Arnott, S., Chandrasekaran, R., Hall, I. H., and Puigjaner, L. C. (1983). Heteronomous DNA. *Nucleic Acids Res.*, **11**, 4141–4155.
- Arnott, S., Chandrasekaran, R., Hall, I. H., Puigjaner, L. C., Walker, J. K., and Wang, M. (1982). DNA secondary structures: helices, wrinkles, and junctions. *Cold Spring Harbor Symp. Quantit. Biol.*, **47**, 53–65.
- Aymami, J., Coll, M., Frederick, C.A., Wang, A. H.-J., and Rich, A. (1989). The propeller DNA conformation of poly(dA).poly(dT). *Nucleic Acids Res.*, **17**, 3229–3245.

- Ban, C., Ramakrishnan, B., and Sundaralingam, M. (1996). Crystal structure of the self-complementary 5'-purine start decamer d(GCGCGCGCGC) in the Z-DNA conformation. I. *Biophysical J.*, **7**, 1215–1221.
- Barbič, A., Zimmer, D. P., and Crothers, D. M. (2003). Structural origins of adenine-tract bending. *Proc. Natl. Acad. Sci. USA*, **100**, 2369–2373.
- Basham, B., Eichman, B. F., and Ho, P. S. (1999). In *Oxford Handbook of Nucleic Acid Structure* (Ed. Neidle, S.), p. 199, Oxford University Press, Oxford.
- Bingman, C. A., Zon, G., and Sundaralingam, M. (1992). Crystal and molecular structure of the A-DNA dodecamer d(CCGTACGTACGG). Choice of fragment helical axis. *J. Mol. Biol.*, **227**, 738–756.
- Calladine, C. R. (1982). Mechanics of sequence-dependent stacking of bases in B-DNA. *J. Mol. Biol.*, **161**, 343–352.
- Calladine, C. R. and Drew, H. R. (1984). A base-centred explanation of the B-to-A transition in DNA. *J. Mol. Biol.*, **178**, 773–782.
- Campos, J. L., Urpi, L., Sanmartin, T., Gouyette, C., and Subirana, J. A. (2005). DNA coiled coils. *Proc. Natl. Acad. Sci. USA*, **102**, 3663–3666.
- Campos, J. L., Valls, N., Urpi, L., Gouyette, C., Sanmartin, T., Richter, M., Alechaga, E., Santaolalla, A., Baldini, R., Creixell, M., Ciurans, R., Skokan, R., Pous, J., and Subirana, J. A. (2006). Overview of the structure of all-AT oligonucleotides: organization in helices and packing interactions. *Biophysical J.*, **91**, 892–903.
- Chandrasekaran, R. and Radha, A. (1992). Structure of poly d(A).poly d(T). *J. Biomol. Struct. Dynamics*, **10**, 153–168.
- Chatake, T., Tanaka, I., Umino, H., Arai, S., and Niimura, N. (2005). The hydration structure of a Z-DNA hexameric duplex determined by a neutron diffraction technique. *Acta Crystallogr.*, **D61**, 1088–1098.
- Chen, Y. Q., Ghosh, S., and Ghosh, G. (1998). A novel DNA recognition mode by the NF-kappa B p65 homodimer. *Nature Struct. Biol.*, **5**, 67–73.
- Chiu, T. K. and Dickerson, R. E. (2000). A crystal structures of B-DNA reveal sequence-specific binding and groove-specific bending of DNA by magnesium and calcium. *J. Mol. Biol.*, **301**, 915–945.
- Clark, G. R., Brown, D. G., Sanderson, M. R., Chwalinski, T., Neidle, S., Veal, J. M., Jones, R. L., Wilson, W. D., Zon, G., Garman, E., and Stuart, D. I. (1990). Crystal and solution structures of the oligonucleotide d(ATGCGCAT)₂: a combined X-ray and NMR study. *Nucleic Acids Res.*, **18**, 5521–5528.
- Coté, M. L., Pflomm, M., and Georgiadis, M. M. (2003). Staying straight with A-tracts: a DNA analog of the HIV-1 polypurine tract. *J. Mol. Biol.*, **330**, 57–74.
- Coté, M. L., Yohannan, S. J., and Georgiadis, M. M. (2000). Use of an N-terminal fragment from moloney murine leukemia virus reverse transcriptase to facilitate crystallization and analysis of a pseudo-16-mer DNA molecule containing G-A mispairs. *Acta Crystallogr.*, **D56**, 1120–1131.
- Crothers, D. M. and Shakked, Z. (1999). In *Oxford Handbook of Nucleic Acid Structure* (Ed. Neidle, S.), p. 455, Oxford University Press, Oxford.
- Crothers, D. M., Haran, T. E., and Nadeau, J. G. (1990). Intrinsically bent DNA. *J. Biol. Chem.*, **265**, 7093–7096.
- Dickerson, R. E. *et al.* (1989). Definitions and nomenclature of nucleic acid structure parameters. *EMBO J.*, **8**, 1–4.
- Dickerson, R. E. (1998). DNA bending: the prevalence of kinkiness and the virtues of normality. *Nucleic Acids Res.*, **26**, 1906–1926.
- Dickerson, R. E. and Drew, H. R. (1981). Structure of a B-DNA dodecamer: II. Influence of base sequence on helix structure. *J. Mol. Biol.*, **149**, 761–786.
- Dickerson, R. E., Goodsell, D., and Kopka, M. L. (1996). MPD and DNA bending in crystals and in solution. *J. Mol. Biol.*, **256**, 108–125.
- Dickerson, R. E., Goodsell, D. S., and Neidle, S. (1994). "...the tyranny of the lattice...". *Proc. Natl. Acad. Sci. USA*, **91**, 3579–3583.
- Diekmann, S., Mazzaralli, J. M., McLaughlin, L. W., von Kitzing, E., and Travers, A. A. (1992). DNA curvature does not require bifurcated hydrogen bonds or pyrimidine methyl groups. *J. Mol. Biol.*, **225**, 729–738.
- DiGabriele, A. D. and Steitz, T. A. (1993). A DNA dodecamer containing an adenine tract crystallizes in a unique lattice and exhibits a new bend. *J. Mol. Biol.*, **231**, 1024–1039.
- Dohm, J. A., Hsu, M.-H., Hwu, J.-R., Huang, R. C. C., Moudrianakis, E. N., Lattman, E. E., and Gittis, A. G. (2005). Influence of ions, hydration, and the transcriptional inhibitor P4N on the conformations of the Sp1 binding site. *J. Mol. Biol.*, **349**, 731–744.
- Drew, H. R. and Dickerson, R. E. (1981). Structure of a B-DNA dodecamer: III. geometry of hydration. *J. Mol. Biol.*, **151**, 535–556.
- Drew, H. R. and Travers, A. A. (1984). DNA structural variations in the E. coli tyrT promoter. *Cell*, **37**, 491–502.
- Edwards, K. J., Brown, D. G., Spink, N., Skelly, J. V., and Neidle, S. (1992). Molecular structure of the B-DNA dodecamer d(CGCAATTTGCG)₂. An examination of propeller twist and minor-groove water structure at 2•2Å resolution. *J. Mol. Biol.*, **226**, 1161–1173.

- Egli, M., Tereshko, V., Teplova, M., Minasov, G., Joachimiak, A., Sanishvili, R., Weeks, C. M., Miller, R., Maier, M. A., An, H., Cook, P. D., and Manoharan, M. (1998). X-ray crystallographic analysis of the hydration of A- and B-form DNA at atomic resolution. *Biopolymers*, **48**, 234–252.
- Egli, M., Williams, L. D., Gao, Q., and Rich, A. (1991). Structure of the pure-spermine form of Z-DNA (magnesium free) at 1-Å resolution. *Biochemistry*, **30**, 11388–11402.
- Eisenstein, M and Shakked, Z. (1995). Hydration patterns and intermolecular interactions in A-DNA crystal structures: Implications for DNA recognition. *J. Mol. Biol.*, **248**, 662–678.
- Eisenstein, M., Frolov, F., Shakked, Z., and Rabinovich, D. (1990). The structure and hydration of the A-DNA fragment d(GGGTACCC) at room temperature and low temperature. *Nucleic Acids Res.*, **18**, 3185–3194.
- Elsawy, K. M., Hodgson, M. K., and Caves, L. S. D. (2005) The physical determinants of the DNA conformational landscape: an analysis of the potential energy surface of single-strand dinucleotides in the conformational space of duplex DNA. *Nucleic Acids Res.*, **33**, 5749–5762.
- Fairall, L., Martin, S., and Rhodes, D. (1989). The DNA binding site of the Xenopus transcription factor IIIA has a non- B-form structure. *EMBO J.*, **8**, 1809–1817.
- Farwar, J., Packer, M. J., and Hunter, C. A. (2006) Prediction of atomic structure from sequence for double helical DNA oligomers. *Biopolymers*, **81**, 51–61.
- Finley, J. B. and Luo, M. (1998). X-ray crystal structures of half the human papilloma virus E2 binding site: d(GACCGCGGTC). *Nucleic Acids Res.*, **26**, 5719–5727.
- Frederick, C.A., Quigley, G.J., Teng, M.-K., Coll, M., van der Marel, G., van Boom, J. H., Rich, A., and Wang, A. H.-J. (1989). Molecular structure of an A-DNA decamer d(ACCGGCCGGT). *Eur. J. Biochem.*, **181**, 295–307.
- Fritsch, V. and Westhof, E. (1991). Three-center hydrogen bonds in DNA: molecular dynamics of poly(dA).poly(dT). *J. Amer. Chem. Soc.*, **113**, 8271–8277.
- Fujii, S., Wang, A. H.-J., Quigley, G. J., Westerink, H., van der Marel, G., van Boom, J. H., and Rich, A. (1985). The octamers d(CGCGCGCG) and d(CGCATGCG) both crystallize as Z-DNA in the same hexagonal lattice. *Biopolymers*, **24**, 243–250.
- Ganunis, R., Guo, H., and Tullius, T. D. (1996). Effect of the crystallizing agent 2-methyl-2,4-pentandiol on the structure of adenine tract DNA in solution. *Biochemistry*, **35**, 13729–13732.
- Gao, Y.-G., Robinson, H., and Wang, A.H.-J (1999). High-resolution A-DNA crystal structures of d(AGGGGCCCCCT). An A-DNA model of poly(dG) x poly(dC). *Eur. J. Biochem.*, **261**, 413–420.
- Gardiner, E. J., Hunter, C. A., Packer, M. J., Palmer, D. S., and Willett, P. (2003). Sequence-dependent DNA structure: a database of octamer structural parameters. *J. Mol. Biol.*, **332**, 1025–1035.
- Gardiner, E. J., Hunter, C. A., Lu, X. J., and Willett, P. (2004). A structural similarity analysis of double-helical DNA. *J. Mol. Biol.*, **343**, 879–889.
- Goodsell, D. S. and Dickerson, R. E. (1994). Bending and curvature calculations in B-DNA. *Nucleic Acids Res.*, **22**, 5497–5503.
- Goodsell, D. S., Grzeskowiak, K., and Dickerson, R. E. (1995). Crystal structure of C-T-C-T-C-G-A-G-A-G. Implications for the structure of the Holliday junction. *Biochemistry*, **34**, 1022–1029.
- Goodsell, D. S., Kaczor-Grzeskowiak, M., and Dickerson, R. E. (1994). Crystal structure of C-C-A-T-T-A-A-T-G-G: implications for bending of B-DNA at A-T steps. *J. Mol. Biol.*, **239**, 79–86.
- Goodsell, D. S., Kopka, M. L., Cascio, D., and Dickerson, R. E. (1993). Crystal structure of CATGGCCATG and its implications for A-tract bending models. *Proc. Natl. Acad. Sci. USA*, **90**, 2930–2934.
- Gorin, A. A., Zhurkin, V. B., and Olson, W. K. (1995). B-DNA twisting correlates with base-pair morphology. *J. Mol. Biol.*, **247**, 34–48.
- Ha, S. C., Lokanath, N. K., Van Quyen, D., Wu, C. A., Lowenhaupt, K., Rich, A., Kim, Y. G., and Kim, K. K. (2004). A poxvirus protein forms a complex with left-handed Z-DNA: crystal structure of a Yatapoxvirus Zalpha bound to DNA. *Proc. Natl. Acad. Sci. USA*, **101**, 14367–14372.
- Ha, S. C., Lowenhaupt, K., Rich, A., Kim, Y. G., and Kim, K. K. (2005). Crystal structure of a junction between B-DNA and Z-DNA reveals two extruded bases. *Nature*, **437**, 1183–1186.
- Hagerman, P. J. (1990). Sequence-directed curvature of DNA. *Ann. Rev. Biochem.*, **59**, 755–781.
- Hamelberg, D., McFail-Isom, L., Williams, L. D., and Wilson, W. D. (2000). Flexible Structure of DNA: Ion dependence of minor-groove structure and dynamics. *J. Amer. Chem. Soc.*, **122**, 10513–10520.
- Hamelberg, D., Williams, L. D., and Wilson, W. D. (2001). Influence of the dynamic positions of cations on the structure of the DNA minor groove: sequence-dependent effects. *J. Amer. Chem. Soc.*, **123**, 7745–7755.
- Hartmann, B., Piazzola, D., and Lavery, R. (1993). BI-BII transitions in B-DNA. *Nucleic Acids Res.*, **21**, 561–568.
- Hays, F. A., Teegarden, A., Jones, Z. J. R., Harms, M., Raup, D., Watson, J., Cavaliere, E., and Ho, P. S. (2005) How sequence defines structure: a crystallographic map of DNA structure and conformation. *Proc. Natl. Acad. Sci. USA*, **102**, 7157–7162.

- Heinemann, U., Alings, C., and Bansal, M. (1992). Double helix conformation, groove dimensions and ligand binding potential of a G/C stretch in B-DNA. *EMBO J.*, **11**, 1931–1939.
- Hizver, J., Rozenberg, H., Frolow, F., Rabinovich, D., and Shakked, Z. (2001). DNA bending by an adenine-thymine tract and its role in gene regulation. *Proc. Natl. Acad. Sci. USA*, **98**, 8490–8495.
- Huang, D.-B., Phelps, C. B., Fusco, A. J., and Ghosh, G. (2005). Crystal structure of a free kappaB DNA: insights into DNA recognition by transcription factor NF-kappaB. *J. Mol. Biol.*, **346**, 147–160.
- Hud, N. V., Sklenář, and Feigon, J. (1999). Localization of ammonium ions in the minor groove of DNA duplexes in solution and the origin of DNA A-tract bending. *J. Mol. Biol.*, **286**, 651–660.
- Hunter, C. A. (1993). Sequence-dependent DNA structure: the role of base stacking interactions. *J. Mol. Biol.*, **230**, 1025–1054.
- Hunter, C. A. and Lu, X.-J. (1997). DNA base-stacking interactions: a comparison of theoretical calculations with oligonucleotide X-ray crystal structures. *J. Mol. Biol.*, **265**, 603–619.
- Jain, S., Zon, G., and Sundaralingam, M. (1991). Hexagonal crystal structure of the A-DNA octamer d(GTGTACAC) and its comparison with the tetragonal structure: correlated variations in helical parameters. *Biochemistry*, **30**, 3567–3576.
- Johansson, E., Parkinson, G., and Neidle, S. (2000). A new crystal form for the dodecamer C-G-C-G-A-A-T-T-C-G-C-G: symmetry effects on sequence-dependent DNA structure. *J. Mol. Biol.*, **300**, 551–561.
- Kielkopf, C. L., Ding, S., Kuhn, P., and Rees, D. C. (2000). Conformational flexibility of B-DNA at 0.74 Å resolution: d(CCAGTACTGG)₂. *J. Mol. Biol.*, **296**, 787–801.
- Koo, H.-S., Drak, J., Rice, J. A., and Crothers, D. M. (1990). Determination of the extent of DNA bending by an adenine-thymine tract. *Biochemistry*, **29**, 4227–4234.
- Kubinec, M. G. and Wemmer, D. E. (1992). NMR evidence for DNA bound water in solution. *J. Amer. Chem. Soc.*, **114**, 8739–8740.
- Lauble, H., Frank, R., Blöcker, H., and Heinemann, U. (1988). Three-dimensional structure of d(GGGATCCC) in the crystalline state. *Nucleic Acids Res.*, **16**, 7799–7816.
- Lipmanov, A., Kopka, M. L., Kaczor-Grzeskowiak, Quintana, J., and Dickerson, R. E. (1993). Structure of the B-DNA decamer C-C-A-A-C-I-T-T-G-G in two different space groups: conformational flexibility of B-DNA. *Biochemistry*, **32**, 1373–1389.
- Liu, J. and Subirana, J. A. (1999). Structure of d(CGCGAATTCGCG) in the presence of Ca(2+) ions. *J. Biol. Chem.*, **274**, 24749–24752.
- Lomonosoff, G. P., Butler, P. J. G., and Klug, A. (1981). Sequence-dependent variation in the conformation of DNA. *J. Mol. Biol.*, **149**, 745–760.
- Luisi, B., Orozco, M., Sponer, J., Luque, F. J., and Shakked, Z. (1998). On the potential role of the amino nitrogen atom as a hydrogen bond acceptor in macromolecules. *J. Mol. Biol.*, **279**, 1123–1136.
- MacDonald, D., Herbert, K., Zhang, X., Polgruto, T., and Lu, P. (2001). Solution structure of an A-tract DNA bend. *J. Mol. Biol.*, **301**, 1081–1098.
- Madhumalar, A. and Bansal, M. (2005). Sequence preference for BI/BII conformations in DNA: MD and crystal structure data analysis. *J. Biomol. Struct. Dynamics*, **23**, 13–27.
- Mahendrasingam, A., Forsyth, V. T., Hussain, R., Greenall, R. J., Pigram, W. J., and Fuller, W. (1986). Time-resolved X-ray diffraction studies of the B in equilibrium D structural transition in the DNA double helix. *Science*, **233**, 195–197.
- Malinina, L., Fernandez, L. G., Huynh-Dinh, T., and Subirana, J. A. (1999). Structure of the d(CGCCGCGGGCG) dodecamer: a kinked A-DNA molecule showing some B-DNA features. *J. Mol. Biol.*, **285**, 1679–1690.
- Marini, J. C., Levene, S. D., Crothers, D. M., and Englund, P. T. (1982). Bent helical structure in kinetoplast DNA. *Proc. Natl. Acad. Sci. USA*, **79**, 7664–7668.
- McCall, M., Brown, T., and Kennard, O. (1985). The crystal structure of d(G-G-G-G-C-C-C-C) a model for poly(dG) poly(dC). *J. Mol. Biol.*, **183**, 385–396.
- McCall, M., Brown, T., Hunter, W. N., and Kennard, O. (1986). The crystal structure of d(G-G-A-T-G-G-G-A-G) forms an essential part of the binding site for transcription factor IIIA. *Nature*, **322**, 661–664.
- Minasov, G., Tereshko, V., and Egli, M. (1999). Atomic-resolution crystal structures of B-DNA reveal specific influences of divalent metal ions on conformation and packing. *J. Mol. Biol.*, **291**, 83–99.
- Montaño, S. P., Coté, M. L., Roth, M. J., and Georgiadis, M. M. (2006). Crystal structures of oligonucleotides including the integrase processing site of the Moloney murine leukemia virus. *Nucleic Acids Res.*, **34**, 5353–5360.
- Nadeau, J. G. and Crothers, D. M. (1989). Structural basis for DNA bending. *Proc. Natl. Acad. Sci. USA*, **86**, 2622–2626.
- Nelson, H. C. M., Finch, J. T., Luisi, B. F., and Klug, A. (1987). The structure of an oligo(dA)-oligo(dT) tract and its biological implications. *Nature*, **330**, 221–226.
- Ng, H.-L. and Dickerson, R. E. (2001). DNA structure from A to B. *Proc. Natl. Acad. Sci. USA*, **98**, 6986–6989.

- Ng, H.-L. and Dickerson, R. E. (2002). Mediation of the A/B-DNA helix transition by G-tracts in the crystal structure of duplex CATGGGCCCATG. *Nucleic Acids Res.*, **30**, 4061–4067.
- Ng, H.-L., Kopka, M. L., and Dickerson, R. E. (2000). The structure of a stable intermediate in the A $\leftrightarrow$ B DNA helix transition. *Proc. Natl. Acad. Sci. USA*, **97**, 2035–2039.
- Olson, W. K., Bansal, M., Burley, S. K., Dickerson, R. E., Gerstein, M., Harvey, S. C., Heinemann, U., Lu, X.-J., Neidle, S., Shakked, Z., Sklenar, H., Suzuki, M., Tung, C.-S., Westhof, E., Wolberger, C., and Berman, H. M. (2001). A standard reference frame for the description of nucleic acid base-pair geometry. *J. Mol. Biol.*, **313**, 229–237.
- Packer, M. J., Dauncey, M. P., and Hunter, C. A. (2000). Sequence-dependent DNA structure: tetranucleotide conformational maps. *J. Mol. Biol.*, **295**, 85–103.
- Packer, M. J. and Hunter, C. A. (1998). Sequence-dependent DNA structure: the role of the sugar-phosphate backbone. *J. Mol. Biol.*, **280**, 407–420.
- Packer, M. J. and Hunter, C. A. (2001). Sequence-Structure Relationships in DNA Oligomers: A Computational Approach. *J. Amer. Chem. Soc.*, **123**, 7399–7406.
- Pohl, F. M. and Jovin, T. M. (1972). Salt-induced co-operative conformational change of a synthetic DNA: equilibrium and kinetic studies with poly (dG-dC). *J. Mol. Biol.*, **67**, 375–396.
- Quintana, J., Grzeskowiak, K., Yanagi, K., and Dickerson, R. E. (1992). Structure of a B-DNA decamer with a central T-A step: C-G-A-T-T-A-A-T-C-G. *J. Mol. Biol.*, **225**, 379–395.
- Rahmouni, A. R. and Wells, R. D. (1992). Direct evidence for the effect of transcription on local DNA supercoiling in vivo. *J. Mol. Biol.*, **223**, 131–144.
- Rhodes, D. and Klug, A. (1981). Sequence-dependent helical periodicity of DNA. *Nature*, **292**, 378–380.
- Rich, A., Nordheim, A., and Wang, A. H.-J. (1984). The chemistry and biology of left-handed Z-DNA. *Ann. Rev. Biochem.*, **53**, 791–846.
- Rich, A. and Zhang, S. (2003). Timeline: Z-DNA: the long road to biological function. *Nature Rev. Genet.*, **4**, 566–572.
- Rohs, R., Sklenar, H., and Shakked, Z. (2005) Structural and energetic origins of sequence-specific DNA bending: Monte Carlo simulations of papillomavirus E2-DNA binding sites. *Structure*, **13**, 1499–1509.
- Rueda, M., Cubero, E., Laughton, C. A., and Orozco, M. (2004). Exploring the counterion atmosphere around DNA: what can be learned from molecular dynamics simulations? *Biophys. J.*, **87**, 800–811.
- Schneider, B., Ginell, S. L., Jones, R., Gaffney, B., and Berman, H. M. (1992). Crystal and molecular structure of a DNA fragment containing a 2-aminoadenine modification: the relationship between conformation, packing, and hydration in Z-DNA hexamers. *Biochemistry*, **31**, 9622–9628.
- Schneider, B., Neidle, S., and Berman, H. M. (1997). Conformations of the sugar-phosphate backbone in helical DNA crystal structures. *Biopolymers*, **42**, 113–124.
- Schuerman, G. S. and Van Meervelt, L. (2000). Conformational flexibility of the DNA backbone. *J. Amer. Chem. Soc.*, **122**, 232–240.
- Schwartz, T., Behike, J., Lowenhaupt, K., Heinemann, U., and Rich, A. (2001). Structure of the DLM-1-Z-DNA complex reveals a conserved family of Z-DNA-binding proteins. *Nature Struct. Biol.*, **8**, 761–765.
- Schwartz, T., Rould, M. A., Lowenhaupt, K., Herbert, A., and Rich, A. (1999). Crystal structure of the Za domain of the human editing enzyme ADAR1 bound to left-handed Z-DNA. *Science*, **284**, 1841–1845.
- Schwieters, C. D. and Clore, G. M. (2007). A physical picture of atomic motions within the Dickerson DNA dodecamer in solution derived from joint ensemble refinement against NMR and large-angle X-ray scattering data. *Biochemistry*, **46**, 1152–1166.
- Shakked, Z., Guerin-Guzikevich, G., Eisenstein, M., Frolow, F., and Rabinovich, D. (1989). The conformation of the DNA double helix in the crystal is dependent on its environment. *Nature*, **342**, 456–460.
- Shatzky-Schwartz, M., Arbuckle, N. D., Eisenstein, M., Rabinovich, D., Bareket-Samish, A., Haran, T. E., Luisi, B. F., and Shakked, Z. (1997). X-ray and solution studies of DNA oligomers and implications for the structural basis of A-tract-dependent curvature. *J. Mol. Biol.*, **267**, 595–623.
- Sherer, E. C., Harris, S. A., Soliva, R., Orozco, M., and Laughton, C. A. (1999). Molecular dynamics studies of DNA A-tract structure and flexibility. *J. Amer. Chem. Soc.*, **121**, 5981–5991.
- Shui, X., McFail-Isom, L., Hu, G. G., and Williams, L. D. (1998). The B-DNA dodecamer at high resolution reveals a spine of water on sodium. *Biochemistry*, **37**, 8341–8355.
- Sims, G. E. and Kim, S.-H. (2003). Global mapping of nucleic acid conformational space: dinucleoside monophosphate conformations and transition pathways among conformational classes. *Nucleic Acids Res.*, **31**, 5607–5616.
- Sprouds, D., Young, M. A., and Beveridge, D. L. (1999). Molecular dynamics studies of axis bending in d(G₅-(GA₄T₄C)₂-C₃) and d(G₅-(GT₄A₄C)₂-C₃): effects of sequence polarity on DNA curvature. *J. Mol. Biol.*, **285**, 1623–1632.
- Sprouds, D., Zacharias, W., Wood, Z. A., and Harvey, S. C. (1995). Dehydrating agents sharply reduce curvature in DNAs containing A tracts. *Nucleic Acids Res.*, **23**, 1816–1821.

- Steffl, R., Wu, H., Ravindranathan, S., Sklenář, V., and Feigon, J. (2004). DNA A-tract bending in three dimensions: solving the dA₄T₄ vs. dT₄A₄ conundrum. *Proc. Natl. Acad. Sci. USA*, **101**, 1177–1182.
- Strahs, D. and Schlick, T. (2000). A-tract bending: insights into experimental structures by computational models. *J. Mol. Biol.*, **301**, 643–663.
- Subirana, J. A. and Faria, T. (1997). Influence of sequence on the conformation of the B-DNA helix. *Biophys. J.*, **73**, 333–338.
- Suzuki, M., Amano, N., Kakinuma, J., and Tateno, M. (1997). Use of a 3D structure data base for understanding sequence-dependent conformational aspects of DNA. *J. Mol. Biol.*, **274**, 421–435.
- Tereshko, V., Minasov, G., and Egli, M. (1999). The Dickerson–Drew B-DNA dodecamer revisited at atomic resolution. *J. Amer. Chem. Soc.*, **121**, 470–471.
- Timsit, Y. and Moras, D. (1991). Groove-backbone interaction in B-DNA: implication for DNA condensation and recombination. *J. Mol. Biol.*, **221**, 919–940.
- Timsit, Y., Vilbois, E., and Moras, D. (1991). Base-pairing shift in the major groove of (CA)_n tracts by B-DNA crystal structures. *Nature*, **354**, 167–170.
- Timsit, Y., Westhof, E., Fuchs, R. P. P., and Moras, D. (1989). Unusual helical packing in crystals of DNA bearing a mutation hot spot. *Nature*, **341**, 459–462.
- Tippin, D. B. and Sundaralingam, M. (1997). Comparison of major groove hydration in isomorphous A-DNA octamers and dependence on base sequence and local helix geometry. *Biochemistry*, **36**, 536–543.
- Trantirek, L., Urbášek, M., Štefl, R., Feigon, J., and Sklenář, V. (2000). A method for direct determination of helical parameters in nucleic acids using residual dipolar couplings. *J. Amer. Chem. Soc.*, **122**, 10454–10455.
- Tullius, T. D. and Dombroski, B. A. (1985). Iron(II) EDTA used to measure the helical twist along any DNA molecule. *Science*, **230**, 679–681.
- Ulanovsky, L. and Trifonov, E. N. (1987). Estimation of wedge components in curved DNA. *Nature*, **326**, 720–722.
- Valls, N., Wright, G., Steiner, R. A., Murshudov, G. N., and Subirana, J. A. (2004). DNA variability in five crystal structures of d(CGCAATTGCG). *Acta Crystallogr.*, **D60**, 680–685.
- Vargason, J. M., Henderson, K., and Ho, P. S. (2001). A crystallographic map of the transition from B-DNA to A-DNA. *Proc. Natl. Acad. Sci. USA*, **98**, 7265–7270.
- Verdaguer, N., Aymami, J., Fernández-Fórner, D., Fita, I., Coll, M., Huynh-Dinh, T., Igolen, J., and Subirana, J. A. (1991). Molecular structure of a complete turn of A-DNA. *J. Mol. Biol.*, **221**, 623–635.
- Vlieghe, D., Turkenberg, J. P., and Van Meervelt, L. (1999). B-DNA at atomic resolution reveals extended hydration patterns. *Acta Crystallogr.*, **D55**, 1495–1502.
- Wahl, M. C. and Sundaralingam, M. (1997). Crystal structures of A-DNA duplexes. *Biopolymers*, **44**, 45–63.
- Wang, A. H.-J., Fujii, A., van Boom, J. H., van der Marel, G. A., van Boeckel, S. A. A., and Rich, A. (1982). Molecular structure of r(GCG)d(TATACGC): a DNA–RNA hybrid helix joined to double helical DNA. *Nature*, **299**, 601–604.
- Wang, A. H.-J., Gessner, R. V., van der Marel, G. A., van Boom, J. H., and Rich, A. (1985). Crystal structure of Z-DNA without an alternating purine–pyrimidine sequence. *Proc. Natl. Acad. Sci. USA*, **82**, 3611–3615.
- Wang, A. H.-J., Hakoshima, T., van der Marel, G., van Boom, J. H., and Rich, A. (1984). AT base pairs are less stable than GC base pairs in Z-DNA: the crystal structure of d(m⁵CGTAm⁵CG). *Cell*, **37**, 321–331.
- Wang, A. H.-J., Quigley, G. J., Kolpak, F. J., Crawford, J. L., van Boom, J. H., van der Marel, G., and Rich, A. (1979). Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature*, **282**, 680–686.
- Wang, A. H.-J., Quigley, G. J., Kolpak, F. J., van der Marel, G., van Boom, J. H., and Rich, A. (1981). Left-handed double helical DNA: variations in the backbone conformation. *Science*, **211**, 171–176.
- Watson, J. D. and Crick, F. H. C. (1953). Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, **171**, 737–738.
- Webster, G. D., Sanderson, M. R., Skelly, J. V., Neidle, S., Swann, P. F., Li, B. F., and Tickle, I. J. (1990). Crystal structure and sequence-dependent conformation of the A.G mispaired oligonucleotide d(CGCAAGCTGGCG). *Proc. Natl. Acad. Sci. USA*, **87**, 6693–6697.
- Wing, R. M., Drew, H. R., Takano, T., Broka, C., Tanaka, S., Itakura, K., and Dickerson, R. E. (1980). Crystal structure analysis of a complete turn of B-DNA. *Nature*, **287**, 755–758.
- Woods, K. K., Machigashi, T., Howerton, S. B., Sines, C. C., Tannenbaum, S. and Williams, L. D. (2004). High-resolution structure of an extended A-tract: [d(CGCAAATTTGCG)]₂. *J. Amer. Chem. Soc.*, **126**, 15330–15331.
- Wu, Z., Delaglio, F., Tjandra, N., Zhurkin, V. B., and Bax, A. (2003). Overall structure and sugar dynamics of a DNA dodecamer from homo- and heteronuclear dipolar couplings and 31P chemical shift anisotropy. *J. Biomol. NMR*, **26**, 297–315.
- Yoon, C., Privé, G. G., Goodsell, D. S., and Dickerson, R. E. (1988). Structure of an alternating-B DNA helix and its relationship to A-tract DNA. *Proc. Natl. Acad. Sci. USA*, **85**, 6332–6336.

- Young, M. A., Ravishanker, G., Beveridge, D. L., and Berman, H. M. (1995). Analysis of local helix bending in crystal structures of DNA oligonucleotides and DNA-protein complexes. *Biophys. J.*, **68**, 2454–2468.
- Yuan, H., Quintana, J., and Dickerson, R. E. (1992). Alternative structures for alternating poly(dA-dT) tracts: the structure of the B-DNA decamer C-G-A-T-A-T-A-T-C-G. *Biochemistry*, **31**, 8009–8021.

Further Reading

- Arnott, S. (1970). The geometry of nucleic acids. *Prog. Biophys. Molec. Biol.*, **6**, 265–319.
- Arnott, S. (2006). DNA polymorphism and the early history of the double helix. *Trends Biochem. Sci.*, **31**, 349–354.
- Berman, H. M. (1997). Crystal studies of B-DNA: The answers and the questions. *Biopolymers*, **44**, 23–44.
- Berman, H. M., Gelbin, A., and Westbrook, J. (1996). Nucleic acid crystallography: a view from the nucleic acid database. *Prog. Biophys. Molec. Biol.*, **66**, 255–288.
- Dickerson, R. E. (1992). DNA structure from A to Z. *Method Enzymol.*, **211**, 67–111.
- Dickerson, R. E. (1999) In *Oxford Handbook of Nucleic Acid Structure* (Ed. Neidle, S.), Oxford University Press, Oxford.
- Egli, M. (2004). Nucleic acid crystallography: current progress. *Curr. Opin. Chem. Biol.*, **8**, 580–591.
- Grzeskowiak, K. (1996). Sequence-dependent structural variation in B-DNA. *Chem. Biol.*, **3**, 785–790.
- Shakked, Z. and Rabinovich D. (1986). The effect of the base sequence on the fine structure of the DNA double helix. *Prog. Biophys. Molec. Biol.*, **47**, 159–195.
- Wahl, M. C. and Sundaralingam, M. (1997). Crystal structures of A-DNA duplexes. *Biopolymers*, **44**, 45–63.

Nonstandard and Higher-Order DNA Structures: DNA–DNA Recognition

4.1 Mismatches in DNA

4.1.1 General Features

Base-pairing in DNA is often considered solely in terms of Watson–Crick hydrogen-bonding, but a large number of other arrangements are possible, and many have been observed experimentally. These are described in this section.

In principle, 16 distinct arrangements for A•T and G•C base-pair hydrogen-bonding are possible, though not all have actually been observed. An important non-Watson–Crick base-pairing pattern was first described by Karst Hoogsteen in 1963 and bears his name. Hoogsteen and reverse Hoogsteen pairs (Fig. 4.1a,b) have been found in several crystal structures of adenine•thymine and adenine•uracil complexes. These arrangements involve atoms N6 and N7 of adenine rather than the N1 and N6 atoms of Watson–Crick hydrogen-bonding – so Hoogsteen hydrogen-bonding with thymine is with the major-groove edge of the adenine base, at right angles from the base edge that forms Watson–Crick pairs in classic A- and B-DNA double helices. Thus Hoogsteen-pairing implies that the adenine base has to be in a *syn* glycosidic angle conformation if the resulting base pair is to be present in an antiparallel DNA duplex.

Mismatched base-pairing in DNA can arise naturally during replication. The possible arrangements are: G•A, G•T, A•A, G•G, T•T, C•C, T•C, or A•C. If left unchecked, these mismatches will lead to mutations and nonsense or incorrect gene products following cellular replication. Cells however have evolved a number of enzymatic repair mechanisms to recognize mispairs, which then attempt to repair them. In general, recognition of a mismatch is followed by duplex unwinding and excision of the mispair. The efficiency of these processes depends on the nature of the mispairing, and often, on sequence context. The details of these mechanisms have been extensively studied in *E. Coli*, and are increasingly well understood in higher organisms. A considerable number of structural studies have been reported on proteins involved in DNA repair, together with their complexes with various damaged DNA sequences. Here we focus on the nature of the initial modified or damaged DNA itself. It has been suggested that many such lesions result in a decrease in base-stacking (Yang, 2006), which itself may be the first recognition signal for initiating repair. It is also likely that

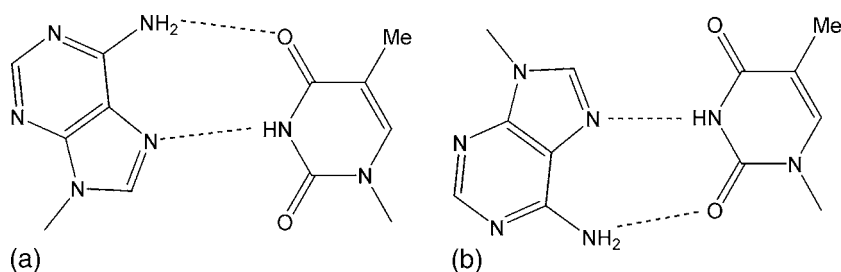


Figure 4.1 Structures of Hoogsteen and reverse Hoogsteen A•T mismatched base pairs.

recognition of local non-B-form DNA is important, perhaps as a bulge or a looped-out base. Sheared base pairs have one base displaced in-plane relative to the other, so that mispairing then occurs (Chou, Chin, and Wang, 2003). These can also be formed during replication or mutation, and may also form in expanded tandem-repetitive DNA sequences in a variety of inherited genetic disorders.

Structural studies on DNA mispairs themselves have addressed the questions of:

- The nature of the hydrogen-bonding involved in the pairings
- How these affect the local and global structure of duplex DNA
- The possible relationships between structure, sequence context of the mispair, and efficiency of mismatch repair

Mismatches generally destabilize duplex DNA. This can be monitored by the extent to which the temperature is decreased at which the transition from a double-helical to an unstructured arrangement occurs. (The transition temperature is termed the melting temperature T_m , and the change is designated as the ΔT_m .) Of the eight mispairs listed above, the A•G one has been the best-studied and is the one that is focused upon here. Structural studies have also been reported on oligonucleotides in A, B, and Z forms with G•T, I(inosine)•T, U•G, I•C and I•A base pairs. The anticancer drug 6-thioguanine (S6G) works by being incorporated into DNA following metabolic conversion to the corresponding nucleotide triphosphate. S6G•C or S6G•T base pairs are closely structurally similar to their unmodified G•C or G•T counterparts (Somerville *et al.*, 2003; Bohon and de los Santos, 2003), albeit significantly less stable (Bohon and de los Santos, 2005).

Mutagenesis (which can eventually lead to carcinogenesis) can occur when external agents such as methylating compounds produce covalent adducts with DNA bases. Structural work on covalently modified oligonucleotides requires milligram quantities of adducts, both for X-ray and NMR studies. This can often be achieved, and a number of structural studies (especially using NMR techniques), have now been performed on oligonucleotides incorporating, for example, methylated bases (Patel, 1992) and metabolites of carcinogens such as benzo[a]pyrene. Of particular interest has been the mutagenic methylation of the O6 atom in guanine, which is produced by a variety of environmental and chemotherapeutic agents.

4.1.2 Purine:Purine Mismatches

Crystal structures have been determined for a number of self-complementary oligonucleotide sequences with A•G base pairs (Table 4.1). These show a variety of

Table 4.1 Some Purine:Purine Mismatched Oligonucleotide Crystal Structures, with the Mismatched Base Pairs Shown in Bold (Distances are in Å)

Sequence	Glycosidic angles	C1'–C1' distances	Reference
5'-CGC GA ATTAGCG GCGATTAA GC GC	G(<i>anti</i>)•A(<i>syn</i>)	10.6, 10.8	Brown <i>et al.</i> , 1988
5'-CGCAAA TTGG CG GCG GT TAAACGC	A(<i>anti</i>)•G(<i>syn</i>)	10.8 (mean)	Leonard, Booth and Brown, 1990
5'-CGCAAGCT GG CG GCG GT CGAACGC	A(<i>syn</i>)•G(<i>anti</i>)	10.7 (mean)	Webster <i>et al.</i> , 1990
5'-CCAAG ATTGG GGTTA GA ACC	A(<i>anti</i>)•G(<i>anti</i>)	12.5	Privé <i>et al.</i> , 1987
5'-CGC GA ATTGGCG GCG GT TAAACGC	G(<i>anti</i>)•G(<i>syn</i>)	10.7, 11.2	Skelly <i>et al.</i> , 1993
5'-CC GA AT GAGG GGAGTAAGCC	A(<i>anti</i>)•G(<i>anti</i>)	8.7	Gao <i>et al.</i> , 1999

base-pairings (Fig. 4.2). The arrangement with both nucleosides in an *anti* conformation necessarily forces the interstrand C1'–C1' distance to be further apart than in a standard B-DNA Watson–Crick base pair (10.4 Å). This *anti, anti* form (Fig. 4.2a) has only been observed in one crystal structure (Leonard, Booth, and Brown, 1990), where there are two consecutive A•G base pairs. Presumably the bulge in the helix produced by their excessive C1'–C1' distances is relieved by the two A•G base pairs being well-stacked on each other. For isolated purine:purine base pairs, it is easier to achieve normal C1'–C1' distances, and hence greater helix stability, by having one base *anti* and the other in a *syn* arrangement (Fig. 4.2b). The base-pairing is of the Hoogsteen type. Energetic considerations suggest that a *syn* conformation would be preferred for the guanine rather than the adenine base, but Table 4.1 shows that this is only sometimes the case.

All A•G-containing DNA oligonucleotides have B-form structures, with generally standard geometries and few distortions from ideality, except that the mispairs tend to be poorly stacked with their neighboring base pairs. Detailed examination of the stacking patterns has shown that the particular glycosidic conformation adopted by a mismatched base is such as to maximize the stacking, and so overrides the slight energy penalty of the adenosine nucleoside being in a *syn* conformation. The nature of the sequence surrounding the mismatch is important, with most of the established structures following the rule that when the mispaired base on the first strand is at the center of the three-residue sequence 5'-Py-**Pu**-Pu, the central purine is then in an *anti* conformation. An oligonucleotide with guanine:guanine base pairs has also been found (Skelly *et al.*, 1993) to have a B-like structure and an *anti, syn* arrangement for the mismatched bases (Fig. 4.3). It appears that this unusual mismatch is associated with B_{II} backbone phosphate conformations (see Chap. 3, Sect. 3.1), which have also been documented for some adjacent G•A mismatches (Chou, Cheng, and Reid, 1992). The B_{II} conformation results in altered phosphate groups intrastrand distances; these changes from standard helical geometry could act as recognition signals for repair nucleases to identify and excise the mispairs at these points.

NMR studies of G•A-containing oligonucleotides have shown that the pattern of glycosidic angles seen in the crystal structures is not always maintained in solution, and some arrangements can be readily converted into others with changes in

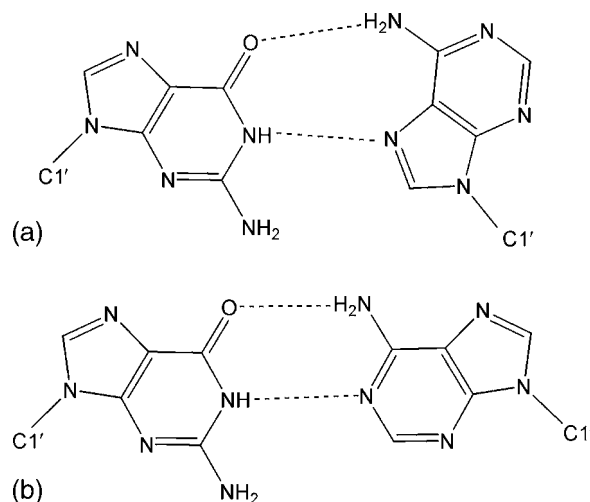


Figure 4.2 Structures of G(*anti*)•A(*anti*) and G(*anti*)•A(*syn*) mismatches.

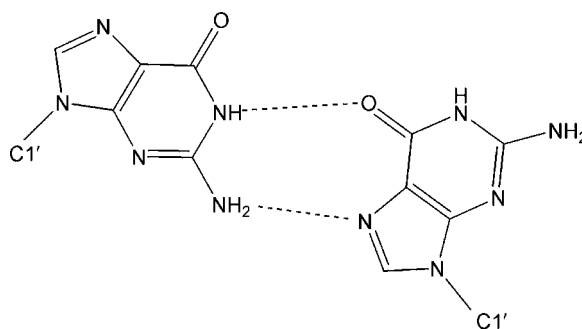


Figure 4.3 A G(*anti*)•G(*syn*) base mismatch, as observed in a mismatch dodecamer crystal structure (Skelly *et al.*, 1993).

pH (Gao and Patel, 1988). Thus, the duplex of d(CGCAAATTGGCG)₂ adopts an A(*anti*)•G(*anti*) arrangement at high pH (Lane *et al.*, 1991), yet an A⁺(*anti*)•G(*syn*) one at low pH. It should be borne in mind that the typical NMR experiment uses somewhat different oligonucleotide solution conditions compared to those in crystallization trials, with higher salt concentrations in the former and high hydrophobic-type solvent concentrations in the latter. These differences are likely to have an effect on the conformational equilibria of equi-energetic mispairs.

The variability of G•A base-pairing arrangements (and hence their stability) is largely dependent on the optimization of base-stacking interactions, and hence on the nature of adjacent bases in a sequence. This is graphically illustrated in the sequence d(ATGAGCGAATA)₂, with four G•A base pairs (Li, Zon, and Wilson, 1991a). NMR and molecular modelling show that these have N7–N2 and N3–N2 hydrogen bonds, with only this arrangement being capable of stabilization by stacking interactions with adjacent bases. Thus, in general, the sequence 5′-Py-**GA**-Pu forms a particularly stable

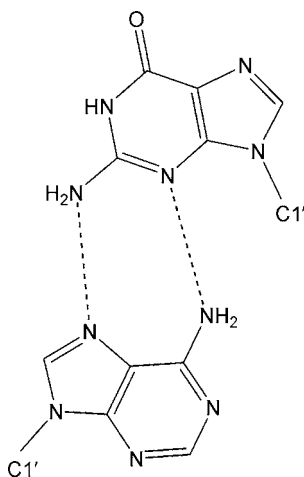


Figure 4.4 The tandem $G(anti)\bullet A(anti)$ mismatch.

mispaired unit (Li, Zon, and Wilson, 1991b), regardless of the details of the hydrogen bonding involved. This stability is seen to be retained in the high-resolution crystal structure (Gao *et al.*, 1999) of the duplex formed by the sequence $d(CCGAATGAGG)_n$, which is a model for centromeric DNA, with the repeated sequence $d(GAATG)_n$. These tandem $G(anti)\bullet A(anti)$ mismatches (Fig. 4.4) in centromeric DNA have been termed “sheared” on account of their side-by-side nature. The notable feature of this crystal structure is the extensive interstrand G/G and A/A base-stacking, which helps to stabilize the unusual G•A base-pairing. These features have also been observed in NMR structures of analogous sequences (Greene *et al.*, 1994; Chou, Zhu, and Reid, 1997). The stability of a particular A•G arrangement is not directly correlated with the kinetics of base excision (the initial step in the repair of this mispair), although it does depend on the sequence context of the mismatch (Sanchez *et al.*, 2003). An A•G mismatch in the sequence $d(GAG)$, with *anti/anti* glycosidic angle conformations had a base-cleavage rate that is threefold faster than with a sequence context of $d(TAG)$, having *syn/anti* glycosidic angle conformations.

4.1.3 Alkylation Mismatches

Methylation at the O6 position of a guanine base in duplex DNA induces the resulting destabilizing $G(OMe)\bullet C$ base pair to undergo a transition mutation to a $G(OMe)\bullet T$ base pair when the DNA is replicated. Methylation of the O6 carbonyl group changes the C6–O6 bond from a double bond to a single one, and hence alters the overall pattern of tautomerism in the guanine ring system, with atom N1 no longer having an attached hydrogen atom. NMR studies on the $G(OMe)\bullet C$ mispair in an oligomer duplex (Kalnik *et al.*, 1989) have indicated that both bases have *anti* glycosidic angles, in an overall B-type helix, with “wobble” hydrogen-bonding solely between N2(G) and O2(C), and with the O6 methyl group oriented toward the cytosine (Fig. 4.5a).

The crystal structure of a (Z-DNA) oligonucleotide duplex containing this mispair (Ginell *et al.*, 1990) shows an arrangement that is more compatible with standard Watson–Crick hydrogen-bonding–distances O6(G)–N4(C) and N1(G)–

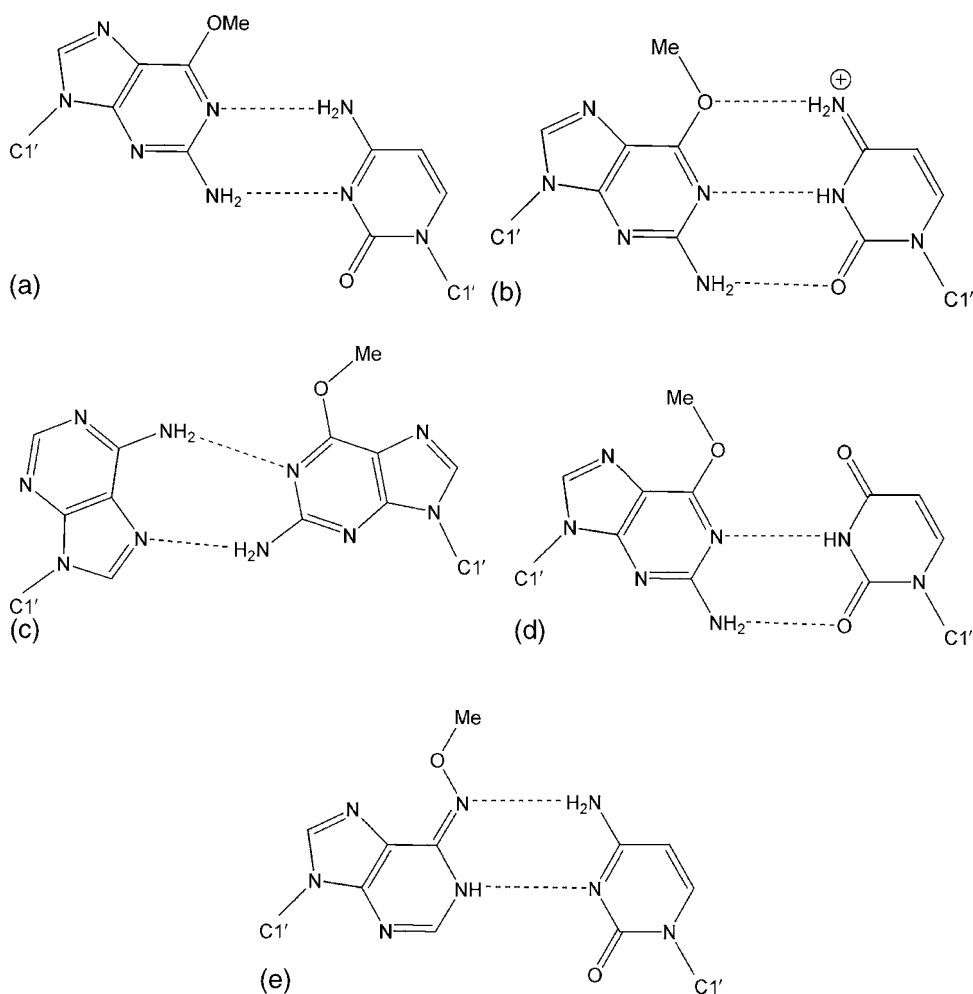


Figure 4.5 (a) The G(OMe)•C wobble base pair. (b) The G(OMe)•C Watson–Crick base pair. (c) The A•G(OMe) base pair. (d) The G(OMe)•T base pair. (e) The A(OMe)•C Watson–Crick base pair.

N3(C) are normal hydrogen-bonding ones (Fig. 4.5b). This suggests that even though the crystals were grown at pH 7.0, atoms N1(G) or N3(C) have acquired a proton so that partial Watson–Crick hydrogen-bonding can occur, which is still less stable than a standard G•C base pair. The various possible arrangements for A•G mismatches have been described in Sect. 4.1.2. When the guanine in this mismatch is methylated, the number of possible hydrogen-bonded base-pair arrangements becomes even greater (Ginell *et al.*, 1994), although sequence context and/or a requirement for adenine protonation will exclude the majority from consideration in specific instances. The crystal structure (Ginell *et al.*, 1994) of a dodecanucleotide duplex with two such A•G(OMe) mismatches, shows an arrangement with two hydrogen bonds and A(*syn*)•G(*anti*) glycosidic angles (Fig. 4.5c). A common theme in this and many other mispair crystal structures is the location of water molecules

that are hydrogen-bonded to (and often bridging between) the bases involved in the mismatches.

The G(OMe)•T base pair, which also destabilizes a duplex, has been observed to have Watson–Crick-type hydrogen-bonding in two B-DNA dodecamer crystal structures (Leonard *et al.*, 1990; Vojtechovsky *et al.*, 1995), with two hydrogen bonds assigned between the two bases (Fig. 4.5d). By contrast, a G(OMe)•T mispair with only a single hydrogen bond, was inferred from an NMR study (Patel *et al.*, 1986) of the sequence d(CGTGAATTCG(OMe)CG), with the mismatch in a distinct sequence context from the crystal structures. The overall shape of the G(OMe)•T base pair in both the NMR and crystallographic structures is close to that of a standard G•C one, whereas the G(OMe)•C one is rather less so, especially in the so-called wobble arrangement (Fig. 4.5a). This difference does not explain why there is only a weak preference for G(OMe)•T base pairs to be incorporated into DNA during replication and not to be readily recognized by repair enzymes.

Mutagens such as hydroxylamine can methoxylate the N6 position of adenine. The crystal structure (Chatake *et al.*, 1999a) of a dodecamer with two methoxyadenine residues shows that these modified bases form analogous Watson–Crick-like base pairs (Fig. 4.5e) with a cytidine on the opposite strand. In order for this to occur, the modified adenines must be in the imino form, as shown, enabling it to mimic a guanine, with the result that cytidine rather than thymine is misincorporated during replication. Thymine opposite this adenine lesion also results in Watson–Crick-like base-pairing (Chatake *et al.*, 1999b).

Oxidative damage to bases can occur by free-radical attack, especially at guanine sites, where the primary product is 7,8-dihydro-8-oxo-guanine. The crystal structure of a duplex containing this lesion (Lipscomb *et al.*, 1995) shows that the standard Watson–Crick G•C hydrogen-bonding arrangements are unimpaired, and alternative G(*syn*) conformations are not involved, in accord with NMR studies (Kouchakdjian *et al.*, 1991). This positions the 8-oxo group in the major groove, where it can be available for recognition by DNA repair enzymes.

A wide range of chemical carcinogens, can bind covalently to DNA, typically as chemically reactive metabolites. These are discussed in Chap. 5, as part of the wider topic of small-molecule DNA recognition (Table 4.2).

Table 4.2 Crystal Structures Incorporating DNA Mismatches

Base pair mismatch	PDB number	NDB number
G•T	113D	BDL009
A•C	1D99	BDL011
G•A	112D	BDL012
A•G	111D	BDL014
A•G	1DMN	BDL022
G•G	1D80	BDL046
A(OMe)•C	456D	BD0009
A(OMe)•T	457D	BD0010
G(OMe)•T	218D	BDLB58
G•A(8-oxo)	1D75	BDLB33
C•G(8-oxo)	183D	BDJB57
G(OMe)•C	1D24	ZDFB21

4.2 DNA Triple Helices

4.2.1 Introduction

The formation of a triple helix by two pyrimidine strands and one purine strand, was discovered (Felsenfeld, Davies, and Rich, 1957) soon after the structure of the double helix itself was determined, when solutions of poly(A) and poly(U) were mixed in appropriate proportions, forming a 1:1:1 three-stranded polynucleotide complex, poly (U•AU). A molecular model for this novel helix was proposed on the basis of fiber diffraction data (Arnott *et al.*, 1976) on both this ribo-polynucleotide and the analogous deoxy-polynucleotide. These studies indicated that the triple helix formed is right-handed. The fiber diffraction data is consistent with an (adenine) purine strand Watson–Crick hydrogen-bonding to a (thymine or uracil) pyrimidine one, and a third (thymine or uracil) pyrimidine strand Hoogsteen hydrogen-bonding to the purine strand and parallel to this strand (Figs. 4.6a and 4.7). This hydrogen-bonding arrangement

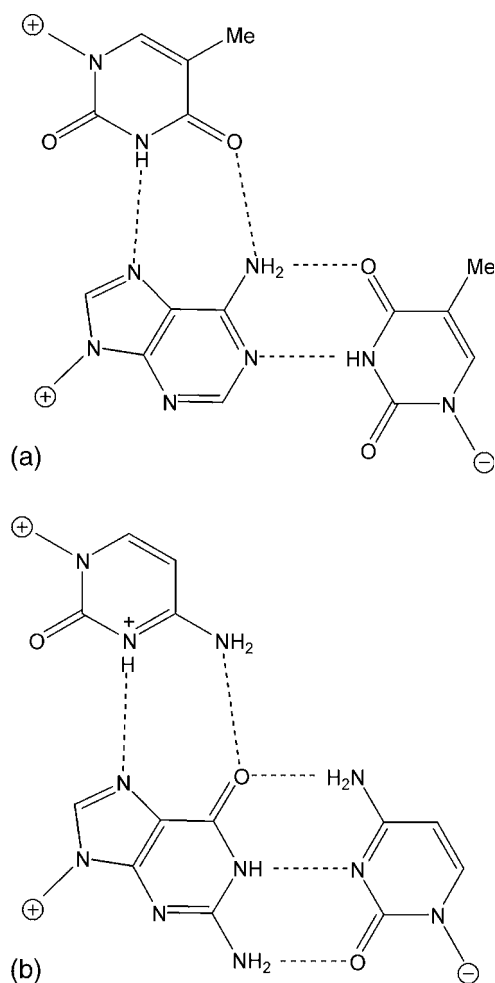


Figure 4.6 The T•AT and C⁺•GC triplex hydrogen-bonding arrangements in a Py-Pu-Py triple helix.



Figure 4.7 Schematic view of an idealized DNA parallel triple helix, with the third (pyrimidine) strand shown shaded.

is termed a base triplet and the triple helix itself is often termed a triplex. We adopt the convention here for the triplet $X \bullet YZ$, that YZ represents the Watson–Crick hydrogen-bonded base pair, with base X being in the third strand and hydrogen-bonded in some way to base Y . Subsequent fiber diffraction studies on both poly ($U \bullet AU$) and its deoxy analogue poly d($T \bullet AT$) have more precisely defined the overall structural parameters of these triple helices (Chandrasekaran, Giacometti, and Arnott, 2000a,b). The 12-fold DNA triple helix has the sugars in all three strands adopting a $C2'$ -*endo* conformation, and has a pitch of 38.4 Å. The poly ($U \bullet AU$) ribonucleotide structure has an 11-fold triplex helix and a more complex set of sugar conformations, with the ribonucleosides in the three strands having $C3'$ -*endo*, $C2'$ -*endo*, and $C2'$ -*endo* puckers respectively. Overall though both triplexes are rather similar, having some features of A-DNA (see below).

A guanine and two cytosine-containing strands can also form a parallel triple helix, which therefore has the same polarity of strands as the $T \bullet AT$ triplex. However, now the third-strand cytosine is required to be protonated at the N3 position in order for two hydrogen bonds to be formed and so enabling effective Hoogsteen hydrogen-bonding to take place (Fig. 4.6b). The $C^+ \bullet GC$ triplex is isosteric with the $T \bullet AT$ one. Protonation of a cytosine base at N3 can only take place at about pH 5.0, well below physiological pH. This places a potentially severe practical limitation on $C^+ \bullet GC$ triplex formation in biological systems, and there has been very

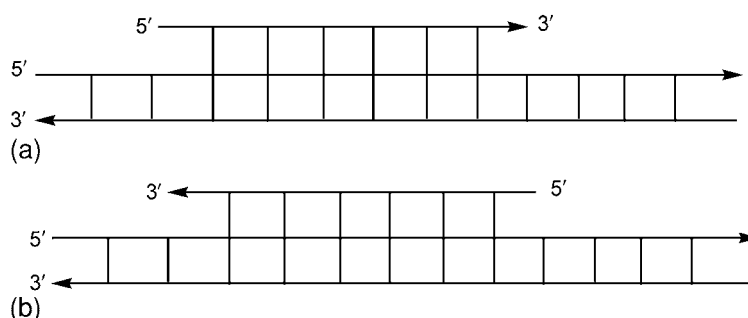


Figure 4.8 Schematic view of the two categories of triple helix formed by an oligonucleotide binding to a long stretch of duplex DNA, with strand directions indicated by arrows. The third strand is shown (a) in a parallel orientation and (b) in an antiparallel orientation to the first duplex strand, which is conventionally a purine one.

considerable effort to circumvent this problem (see below). Substitution of a methyl group or bromine at the 5-position of cytosine results in some stabilization of triplex formation for C⁺•GC-containing sequences at ca pH 7, (i.e. around physiological pH). The precise reasons for this are not clear. It is possible that there is an increased hydrophobic contribution to third-strand binding when there are these substituents at the 5-position of cytosine, rather than there being a shift in the pK_a of the N3 atom of cytosine. Triple helices can be stable under some circumstances with both cytosines and thymines in the third strand, i.e. with mixed pyrimidine sequences. These are further discussed below.

The triple helix phenomenon was for many years no more than a laboratory curiosity, until it was found that stretches of triple helix could be formed by oligonucleotides of appropriate sequence binding to much longer duplex DNA molecules (Fig. 4.8a). For example, the oligonucleotide sequence, (5'-dTTCCTTTCTTCTTTCTTTT), with a covalently attached strand cleavage agent, has been used to bind to and cleave a unique site on a yeast chromosome (Strobel and Dervan, 1990), demonstrating the exceptionally high specificity for a target duplex sequence that triplex formation can achieve. In spite of this specificity, triple helices are inherently readily disrupted as a consequence of the proximity of the third strand and its negatively charged phosphate groups to the two duplex strands.

4.2.2 Structural Studies

The model derived from the original fiber diffraction studies (Arnott *et al.*, 1976) suggested that the third strand in this ribonucleotide triple helix occupies the major groove of the purine–pyrimidine double helix. This key feature has been retained in all subsequent models and ultimately has been verified by both NMR and single-crystal studies. The parallel triple helix in the fiber diffraction model has some (though not all) features of classic A-DNA (Fig. 4.9). It has bases significantly (~ 3.5 Å) displaced from the helix axis and an average helical twist of $\sim 30^\circ$. The duplex minor-groove width of 10.7 Å is almost identical to the A-DNA one of 11.0 Å. However its wide major-groove width, of 9.8 Å, which is necessary to accommodate the third strand, is much greater than in standard duplex A-DNA. All deoxyribose sugars in this model have C3'-*endo* pucker.

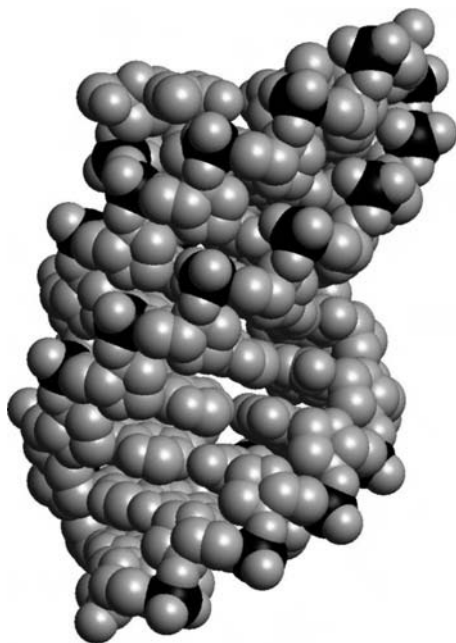


Figure 4.9 The A-form parallel triple helix, as suggested from fiber diffraction studies (Arnott *et al.*, 1976), shown as a space-filling structure. Phosphorus atoms are shaded black, and the third Hoogsteen strand is at the front of this view, nestling in the enlarged major groove.

Triple helices can be formed by relatively short oligonucleotides in solution, such as $d(A)_{12} \cdot 2d(T)_{12}$, and a number of NMR studies have been performed (see, for example, de los Santos, Rosen, and Patel, 1989; Rajagopal and Feigon, 1989; Radhakrishnan *et al.*, 1992; Macaya *et al.*, 1992; Radhakrishnan and Patel, 1994; Pasternack *et al.*, 2002) on such systems. Intramolecular triplexes formed by a single strand of appropriate sequence folding back on itself tend to be more stable and have been especially well-studied. The sugar residues in parallel triplexes tend to have B-DNA type C2'-*endo* sugar puckers, whereas the helical twists are closer to those of A-DNA. Overall the parallel triple helices derived from NMR data have morphology closest to that of B-type duplex DNA, with the differences in, for example, x -displacement, being due to the need to expand the size of the major groove to accommodate the third strand (Fig. 4.7). The third DNA strand in a parallel triplex can be replaced by a RNA one, with some increase in stability. This hybrid triple helix still retains much of its B-like character (Gotfredsen, Schultze, and Feigon, 1998), in spite of the presence of the RNA strand, with sugar puckers of the DNA component in the C2'-*endo* range and an x -displacement of $-0.7 \text{ \AA}$.

A number of molecular mechanics and dynamics simulation studies have been performed on triple-helical DNAs. Until recently these studies have not produced a consistent single structural model, but rather have shown a diversity of structural types, with differing helix morphologies and sugar puckers (Cheng and Pettit, 1992; Laughton and Neidle, 1992). This is a reflection of several factors that have been common to many nucleic acid simulations, but are probably more marked for triplexes because of the challenge of being able to adequately treat the electrostatics of three negatively charged DNA strands in close proximity, which could be addressed more adequately with current simulation methodology, for example, by the use of the

particle-mesh Ewald algorithm. Differences in sugar pucker are in large part a result of the earlier use of different force fields with subtly different pucker preferences. Thus the increasing ability to undertake long-time-scale stable simulations, together with improved electrostatic treatment, has been able to address both issues, and has resulted in simulated parallel triplex B-type structures that have many features in common with experimental data (Shields, Laughton, and Orozco, 1997).

Very few single-crystal triplex structures have been determined, in spite of much effort, which has often produced at best crystals with excellent external morphology but little long-range internal order (Liu *et al.*, 1996). Several crystal structures of duplex DNA oligonucleotides with overhanging ends have been determined. These have revealed the details of triplet arrangements of bases in their crystal lattices. For example, the structure of d(GCGAATTCG) involves Hoogsteen G•GC base triplets formed by interactions between adjacent molecules in the crystal (Van Meervelt *et al.*, 1995). However the stretches of triplets in these structures are too short to provide information about the morphology of triple helices themselves. A significantly longer triple-stranded arrangement has been found in the structure (Betts *et al.*, 1995) of a complex between an 18-mer oligopyrimidine foldback peptide nucleic acid (PNA) sequence, and a 9-mer oligopurine sequence (Fig. 4.10). PNA molecules are nucleic acid mimics in which the sugar-phosphate linkage between successive bases has been replaced by an uncharged 2-aminoethylglycine unit with closely similar spatial properties to the natural linker (Fig. 4.10). They form very stable hybridization

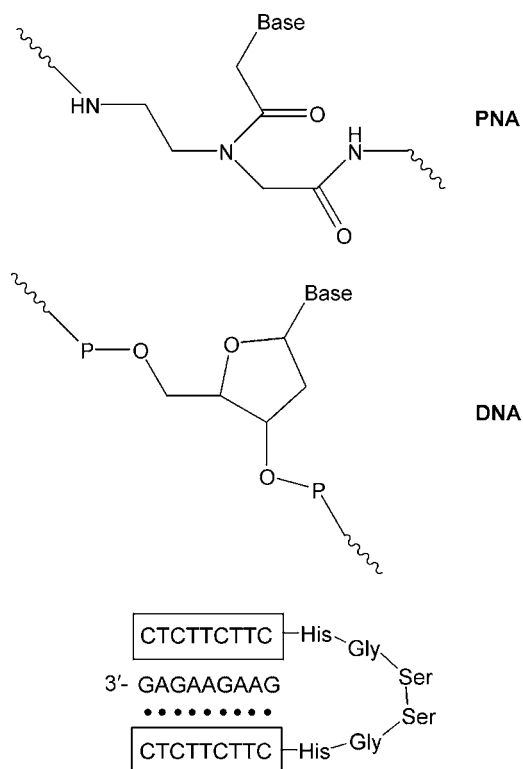


Figure 4.10 The molecular structure of the peptide nucleic acid (PNA) repeating unit, compared to a DNA backbone. Also shown are the sequences of DNA and PNA strands in the crystal structure of the foldback PNA-DNA triple helix (Betts *et al.*, 1995). The PNA strands are shown highlighted in boxes.

complexes both with other PNA molecules and with DNA oligonucleotides (Nielsen *et al.*, 1991; Uhlmann, 1998). The PNA-DNA triplex crystal structure (Fig. 4.11) has features distinct from those of either A- or B-DNA, with 16 bases per turn and a large x -displacement, of 6.8 Å (Betts *et al.*, 1995). All sugars of the DNA strand have C3'-*endo* puckers. All of the ten T•AT and eight C⁺•GC base triplets in this structure have the expected Hoogsteen geometry. The overall morphology of this triplex, which has been termed the P-form, is governed by the particular requirements of the PNA strands, which are known to prefer to adopt A-type geometries. A more recent PNA crystal structure (Petersson *et al.*, 2005), has a complex lattice arrangement of left- and right-handed duplex structures, which interact together to form a triplex.

A true DNA triple helix has been crystallized (Rhee *et al.*, 1999), using a construct of three overlapping oligonucleotide sequences, such that there is a central symmetric segment of six base triplets forming a short stable triplex region, flanked by a hexamer duplex at each end (Fig. 4.12). Crystals were grown at pH 5.0, since the dominance of C⁺•GC triplets in the sequence requires a low pH for optimum stability. The resulting crystal structure (Fig. 4.13), at 1.8 Å resolution, confirms all of the general features of triplexes inferred from earlier modelling and NMR studies. In particular, the short distances (2.7 Å) between N3 of a cytosine and N7 of a guanine in a C⁺•GC triplet confirm the existence of a second hydrogen bond in this C•G Hoogsteen base pair, and thus of cytosine N3 protonation (Fig. 4.14). Most of the sugars have C2'-*endo* pucker. This feature, combined with the low base inclination of 6° in the triplex region and the x -displacement of -2 Å, suggests that the triplex part of the structure could be categorized closer to the B-type family than the A-type. However, as with all triplex

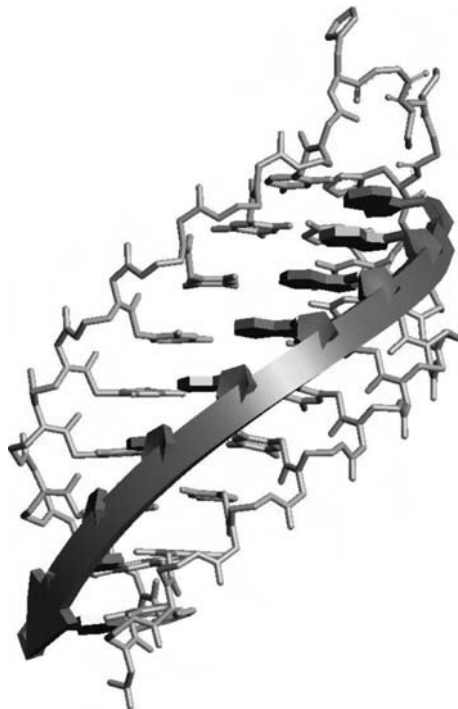


Figure 4.11 The crystal structure of the PNA-DNA triple helix, with the DNA strand highlighted (Betts *et al.*, 1995).



Figure 4.12 The sequences cocrystallized to form a parallel DNA triple helix (Rhee *et al.*, 1999), with the triple helical segment highlighted.



Figure 4.13 The crystal structure of the parallel DNA triple helix (Rhee *et al.*, 1999).

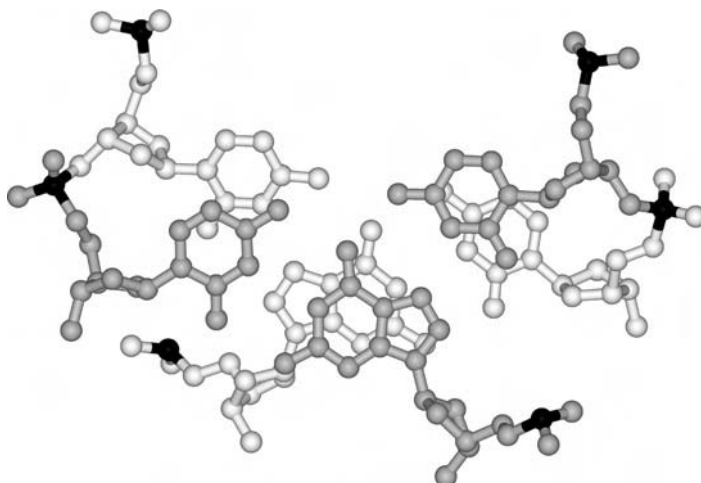


Figure 4.14 View onto two successive base triplets in the triple-helix crystal structure (Rhee *et al.*, 1999).

structures, other features do not accord with this classification, and the structure is best not described in such terms.

Locked nucleic acids (see below) can form stable triple helices, and one such structure has been solved by NMR methods (Sorensen, Nielsen, and Petersen, 2004). The third (pyrimidine) strand has alternating DNA and LNA nucleotides, but with sequences chosen to be identical to those previously studied as standard DNA.DNA•RNA and DNA.DNA•DNA triple helices (Tarköy *et al.*, 1998; Gotfredsen, Schultze, and Feigon, 1998). The Watson–Crick duplex part of the structure is intermediate between A- and B-forms, with a high mean propeller twist of 21°, a mean inclination to the helix axis of 10°, and sugar puckers in the C2'-*endo* region. The sugar puckers in the third strand are alternating between C2'-*endo* and C3'-*endo* forms. The large propeller twists between the Hoogsteen bases, as well as the Watson–Crick ones results in highly unusual (albeit weak) additional hydrogen-bonding between pyrimidines, which form a spiral running along the whole of the structure. However overall the morphology is surprisingly similar to other (natural) triple helices (Fig. 4.15).

4.2.3 Antiparallel Triplexes and Nonstandard Base-pairings

Triple helices can also be formed with the third strand consisting solely of purine residues, so that the resulting helix has Pu•PuPy triplets of bases (Beal and Dervan, 1991). Such a triple helix forms most readily with G•GC base triplets, having a G•G hydrogen-bonding arrangement that is likely to be as shown in Fig. 4.16a, with the third, purine strand being antiparallel to the purine strand of the duplex. This arrangement is pH-independent, contrasting with the C⁺•GC triple helix. The overall stability of antiparallel triplexes is frequently greater than that of parallel ones. The guanosine nucleoside on the third strand of the G•GC base triplet can have either *anti* or *syn* arrangements (Fig. 4.16a, b). Both NMR (Radhakrishnan, de los Santos, and Patel, 1991; Radhakrishnan and Patel, 1994) and molecular simulation studies (Laughton and Neidle, 1992) suggest that the *anti* one is actually preferred, though no crystal structures of antiparallel triplexes have been reported to date (Table 4.3).

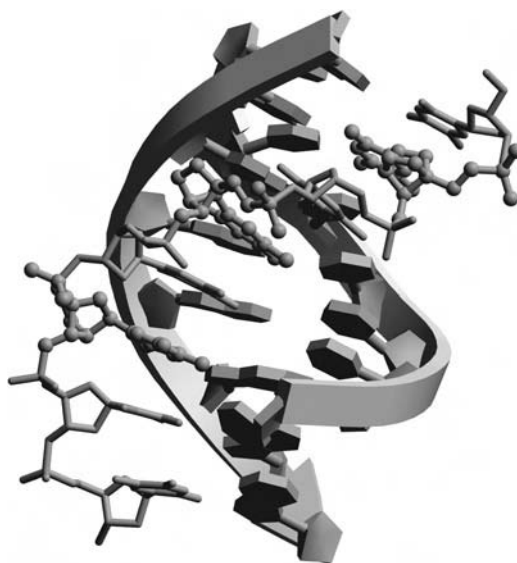


Figure 4.15 One of the NMR structures determined for a locked nucleic acid DNA•DNA-LNA hybrid (Sorensen, Nielsen, and Petersen, 2004). The DNA•DNA duplex is shown as a ribbon structure. The three LNA residues in the third strand are shown in ball-and-stick representation.

There have been numerous attempts to extend the sequence restrictions on triple-strand recognition (Gowers and Fox, 1999) using natural DNA bases. There are two overriding challenges that any such an approach has to overcome:

- The need to maximize base–base triplet hydrogen-bonding
- The distortions in DNA structure that are introduced when inserting one or a few purines in a oligopyrimidine third strand (and vice versa). These distortions reduce intrastrand base-stacking and are probably a major contributor to the instability of many possible base triplets

However, some particular combinations have been shown to stabilize the resulting structures. For example, a thymine base in a pyrimidine third strand can be replaced by a guanine, with the resultant G•TA base triplet (Fig. 4.17a) being relatively stable (Griffin and Dervan, 1989). However the distortions that this is likely to produce in the DNA backbone suggest a limit on the number of such triplets that can be tolerated within a sequence. Almost all other possible triplet mismatches are highly destabilizing to triple-helix formation (Mergny *et al.*, 1991; Kiessling, Griffin, and Dervan, 1992) such that triplex formation for an otherwise stable sequence can be abolished by even a single such mismatch within it. The nature of the flanking sequences is generally important for stability of all types of nonstandard triplet sequence. There are several mismatches, such as T•CG (Fig. 4.17b), that may be stable in some flanking sequence circumstances, but clearly do not work in others. Table 4.4 details the “code” for triple-strand recognition.

More general mixed sequence recognition of all four Watson–Crick base pairs in a duplex (i.e. C•G, G•C, A•T, and T•A) is almost certainly only achievable with nonstandard, unnatural “designer” bases (Griffin *et al.*, 1992). These may be capable of being incorporated into a general-sequence oligonucleotide and hybridizing with a target duplex with high affinity. The use of positively charged analogues of thymine (e.g. 5-(1-propargylamino)-2′-deoxyuridine, Fig. 4.18), results in a 100-increase in binding affinity (Bija-

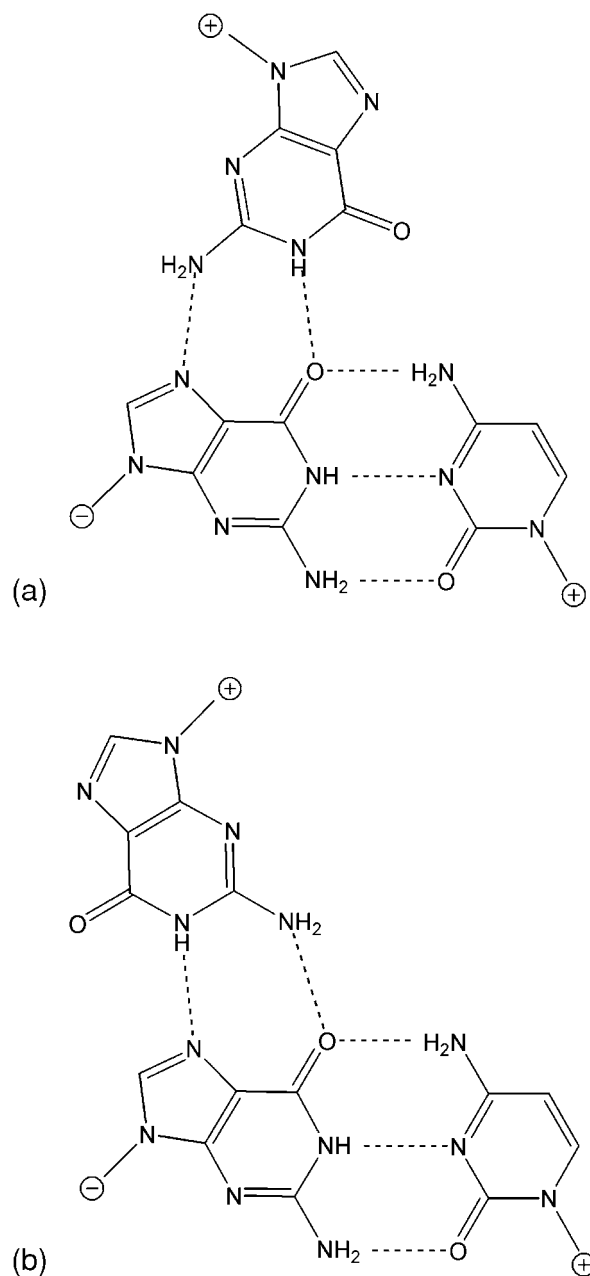


Figure 4.16 Antiparallel and parallel G•GC base triplet hydrogen-bonding arrangements.

Table 4.3 Crystal and NMR Structures of Nucleic Acid Triple Helices

Triple helix type	PDB code No.	NDB code No.
DNA•DNA–DNA	136D	-
PNA•DNA–PNA	1PNN	PNA001
DNA•DNA–DNA	1D3R	BD0017
DNA•DNA–LNA	1W86	-

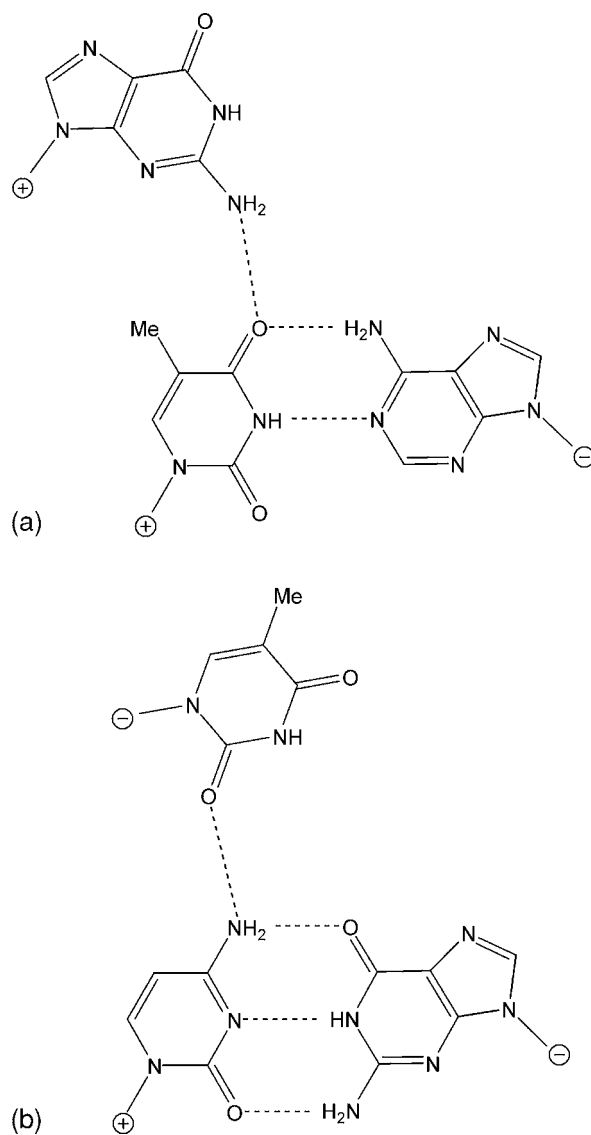


Figure 4.17 (a) The G•TA base triplet. (b) The T•CG base triplet.

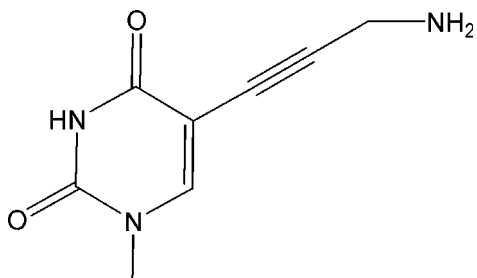


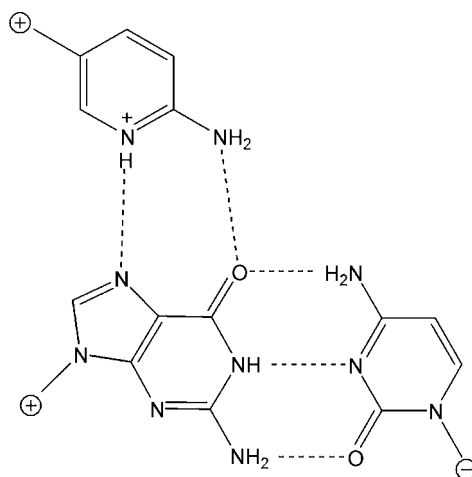
Figure 4.18 The structure of 5-(1-propargylamino)-2'-deoxyuridine (Bijapur *et al.*, 1999).

Table 4.4 Relative Stabilities of All 16 Possible Base Triplets. (Adapted from Soyfer and Potaman, 1996.)

Triplet	Stability	Triplet	Stability
$\text{C}^+\bullet\text{GC}$	++++	$\text{C}\bullet\text{CG}$	+
$\text{A}^+\bullet\text{GC}$	++	$\text{A}\bullet\text{CG}$	–
$\text{G}\bullet\text{GC}$	+	$\text{G}\bullet\text{CG}$	+
$\text{T}\bullet\text{GC}$	++	$\text{T}\bullet\text{CG}$	++
$\text{C}\bullet\text{AT}$	+	$\text{C}\bullet\text{TA}$	–
$\text{A}\bullet\text{AT}$	+	$\text{A}\bullet\text{TA}$	–
$\text{G}\bullet\text{AT}$	+	$\text{G}\bullet\text{TA}$	+++
$\text{T}\bullet\text{AT}$	++++	$\text{T}\bullet\text{TA}$	+

pur *et al.*, 1999). This is possibly as a result of favourable electrostatic interactions with adjacent phosphate groups. Adding a second amino substituent, results in, for example, 2'-aminoethoxy-5-(3-aminoprop-1-ynyl)uridine, which produces a significant further improvement in triplex stability, even at pH 7.0 (Osborne *et al.*, 2004). Active recognition of all four bases in a target duplex can be achieved by extending this concept with a third strand containing four different synthetic nucleotides (Rusling *et al.*, 2005). The TA step in a sequence can be as effectively recognized as the AT step is recognized by T, by using a synthetic thiazolyl-alanine deoxynucleoside (Wang *et al.*, 2005) (Table 4.4).

The requirement for the $\text{C}^+\bullet\text{GC}$ triplet to be protonated is a major limitation when selecting potential genomic targets for triplex recognition (see below). This is especially constraining with contiguous $\text{C}^+\bullet\text{GC}$ triplets. There have been numerous attempts to devise nonnatural cytosine mimics that provide two appropriate hydrogen bonds to the Hoogsteen face of the guanine. One involves the C-nucleoside analogue 2-aminopyridine, which has a pK_a of 6.86 compared to that of 4.3 for deoxycytidine itself, and will thus be largely protonated at neutral pH, so ensuring that two hydrogen bonds are likely to be formed (Fig. 4.19). This has been borne out by the formation of a stable triplex at pH 7 with 2-aminopyridine-containing third strands (Bates *et al.*, 1996; Hildebrand, Parel, and Levman, 1997). As yet none of the many cytosine mimics have progressed beyond the experimental stage to become generally available, in large part because of the high cost of synthetic scale-up.

**Figure 4.19** The triplet formed by 2-aminopyridine with a $\text{G}\bullet\text{C}$ base pair (Bates *et al.*, 1986; Hildebrand *et al.*, 1997).

The relative weakness of the third strand association in triple helices can be enhanced by triplex-selective ligands. These are typically molecules that have the ability to intercalate between base triplets, in a manner analogous to conventional duplex intercalation (see Chap. 5). Molecules such as acridines can be covalently attached to either the 5' or 3' end of the third strand oligonucleotide, and rational design considerations have enabled the optimization of a variety of ligands to be undertaken (Escudé *et al.*, 1998; Keppler, Neidle, and Fox, 2001). Attachment of a ligand to covalently cross-link the target duplex provides yet greater enhancement of binding. Psoralen has been a well-studied bifunctional cross-linking agent, binding at the sequence 5'-TpA, thereby placing a restriction on its general applicability for triplexes. In addition psoralen has a requirement for UV activation in order to form covalent cross-links. An alternative approach is to use backbone-modified oligonucleotides for enhancement of stabilization. The use of N3'-P phosphoramidite-modified oligonucleotides is of especial promise, since they hybridize to duplex sequences with significantly higher affinity than unmodified ones, and they are stable in cellular environments – this backbone modification is resistant to nuclease degradation (Escudé *et al.*, 1996).

4.2.4 Triplex Applications

Much of the interest in the triplex phenomenon arises from the findings that triplex formation using an appropriate oligonucleotide can specifically inhibit transcription of a particular gene (Grigoriev *et al.*, 1993), for example, by binding to a promoter region that contains a suitable triplex-forming sequence. This is termed the “antigene” approach to artificial gene regulation, and complements attempts to read DNA sequences with purely synthetic molecules targeted against the minor groove (see Chap. 5). Both methods require a target site of ca. 16–20 bases for uniqueness in the human genome, which can readily be achieved by a triplex-forming oligonucleotide provided that the target site has an uninterrupted run of purines or pyrimidines. Since the stability of a triplex is length-dependent, an even longer sequence is to be preferred. The (unsurprising) relative rarity of perfect triplex-forming sequences has encouraged the studies that have been outlined above, which are aimed at extending the current sequence restrictions on triplex formation. The goal of such studies is triplex recognition of a regulatory or coding sequence for *any* gene that is a suitable therapeutic target, and thus in principle is not confined to those with perfect oligopyrimidine or oligopurine triplex sites.

Both parallel and antiparallel triplex-forming oligonucleotides have been used in a wide range of *in vitro* and cell-based experiments to demonstrate the viability of the triplex approach, especially its ability to downregulate endogenous genes such as the oncogene *c-myc* (Catapano *et al.*, 2000). This goal has been conclusively demonstrated against an HIV polypurine tract in cell culture (Faria *et al.*, 2000), using N3'-P phosphoramidite-modified oligonucleotides to optimize triplex stability and produce inhibition of transcription elongation. Extension of triplex approach to *in vivo* situations are still at an early stage of development (Giovannangeli and Hélène, 2000), not least because of the difficulties involved in ensuring that oligonucleotides are effectively transported into the cell nucleus.

A web-based search engine is now available from <http://spi.mdanderson.org/tfo/> to search for potential triplex-forming sequences throughout the human and other genomes (Gaddis *et al.*, 2006).

4.3 Guanine Quadruplexes

4.3.1 Introduction

It has long been known that runs of guanosine nucleosides in oligo- and polynucleotides can readily aggregate together, provided a monovalent cation such as potassium or sodium is present to provide stabilization. Diffraction patterns from fibers of poly(G) have been interpreted as arising from a novel four-stranded helix (Gellert, Lipsett, and Davies, 1962; Arnott, Chandrasekaran, and Marttila, 1974; Zimmerman, Cohen, and Davies, 1975), with four guanine bases, one from each strand, being involved in a planar four-stranded (tetrad) arrangement with G–G base-pairing between them (Fig. 4.20). This very stable arrangement is termed a guanine (G)-quartet, or G-tetrad. It involves a total of eight hydrogen bonds between the “Watson–Crick” face of one guanine base and the “Hoogsteen” major-groove face of another, in a manner analogous to that observed in some G•G mismatched duplexes.

There has been increasing interest in nucleic acids containing the G-tetrad motif, not least because guanine-rich sequences have widespread prevalence in all genomes, and have been identified, for example, in promoter and immunoglobulin switch regions and in recombination hot spots (Simonsson, 2001). Their occurrence in the DNA at the ends of eukaryotic chromosomes (known as telomeres), has been especially well-characterized (de Lange, 2005). Telomeric DNA comprises tandem repeats of short guanine tracts, sometimes including adenines/thymines, such as G_nT_n , $G_nT_nG_n$, G_nA_n , or $(TTAGGG)_n$, together with associated telomeric proteins. The nature of the repeat is species-dependent, with $(TTAGGG)_n$ being the repeat in mammals. The function of telomeres is to protect chromosomal ends from unwanted recombination

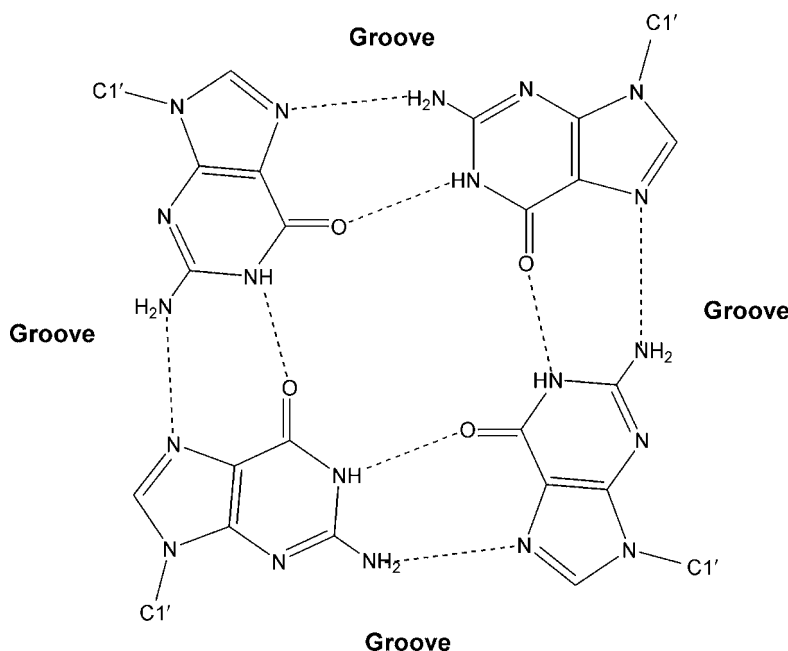
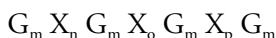


Figure 4.20 The arrangement of hydrogen bonds in the guanine tetrad.

and nuclease damage. Human telomeric DNA in somatic cells is typically 8–10 kb, of duplex DNA. However the terminal 100–200 nucleotides at the 3′ end are single-stranded (Wright *et al.*, 1997), and are therefore in principle conformationally unconstrained, although in cells this single-strand overhang is associated with the single-strand binding protein POT1 (Lei, Podell, and Cech, 2004). Single-stranded telomeric DNA in the absence of protein can fold back and then dimerize to form four-stranded hairpin loops that can be stabilized (Sen and Glibert, 1988; Sundquist and Klug, 1989) by the formation of the guanine tetrad motif outlined above. It can also form intramolecular structures by repeated foldbacks. These G-tetrad-containing structures are termed quadruplexes or tetraplexes (Gilbert and Feigon, 1999), and can also be formed by short-length oligonucleotides of appropriate sequence, typified as:



where m is the number of G residues in each short G-tract. These are usually directly involved in G-tetrad interactions. X_n , X_o and X_p can be any combination of residues, including G, forming loops between the G-tetrads.

Several of these sequences have been shown to possess aptamer properties when folded into quadruplex arrangements, and have been used to target a range of cellular proteins such as thrombin and STAT-3 (e.g., Jing *et al.*, 2006). One particular family of quadruplexes has remarkable anticancer properties (McMicken, Bates, and Chen, 2003), and is currently in clinical trials; their mechanism of action, though not yet fully elucidated, may involve the protein nucleolin and its role in RNA processing.

The evidence for the existence of stable quadruplex structures *in vivo* remains controversial, although a number of quadruplex-binding proteins have been identified and convincing data from specific staining of telomeres with antibodies raised against quadruplex sequences is available (Maizels, 2006). On the other hand there is now considerable data supporting the concept that quadruplex structures can be induced to form, and not only within telomeric sequences. Appropriate G-tract sequences with quadruplex-forming potential have been identified using bioinformatics approaches in many regions within the human genome (notably in promoter sequences), as well as in bacterial and other genomes (Eddy and Maizels, 2006; Huppert and Balasubramanian, 2005, 2007; Todd, Johnston, and Neidle, 2005; Rawal *et al.*, 2006). The occurrence of a particular quadruplex motif sequence does not necessarily indicate that it forms a stable quadruplex structure, either within its genomic environment, or even as an isolated sequence. A standard approach to begin answering this question is to use a combination of biophysical probe methods such as NMR and circular dichroism, and the monitoring of melting behavior (Mergny, Phan, and Lacroix, 1998). Quadruplex-forming promoter sequences have been located in a number of oncogenes and cancer-relevant genes, such as *K-ras* (Cogio and Xodo, 2006), *c-kit* (Rankin *et al.*, 2005) and *bcl2* (Dexheimer, Sun, and Hurley, 2006). Several G-rich sequences that form quadruplexes have been identified in the NHE III nuclease hypersensitive region of the *c-myc* oncogene (Simonsson, 2001; Siddiqui-Jain *et al.*, 2002), and these exceptionally stable quadruplexes have been characterized by biophysical methods, with several NMR structures now being available (see below). It has been proposed that induction of quadruplex structures by quadruplex-specific small molecules could be used to downregulate the expression of such genes, and ligands such as the porphyrin TMPyP4 have been studied for this purpose (Siddiqui-Jain *et al.*, 2002; Hurley *et al.*, 2006). Quadruplex-binding small molecules are discussed further in Chap. 5.

4.3.2 Overall Structural Features of Quadruplex DNA

There are a number of ways of forming a quadruplex structure from a given quadruplex sequence, depending on the number of short G-tracts. Four separate strands can associate together in an intermolecular (unimolecular) manner (Fig. 4.21) or by a single strand folding back on itself in an intramolecular structure (Fig. 4.22a,b). These two topologies differ in the orientations of the phosphodiester chain within the various strands of the complex. Whether the strands are parallel or antiparallel to each other depends in part on the nature and length of the sequence, and sometimes as well as on the counter-ion used.

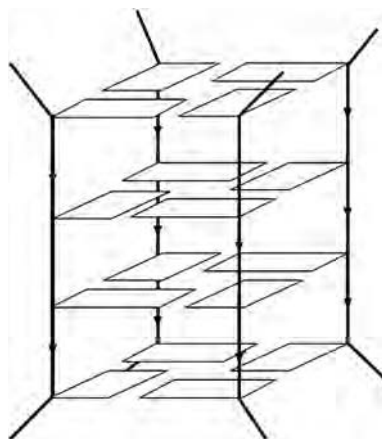


Figure 4.21 Schematic of an all-parallel tetramolecular G-quadruplex with four G-tetrads.

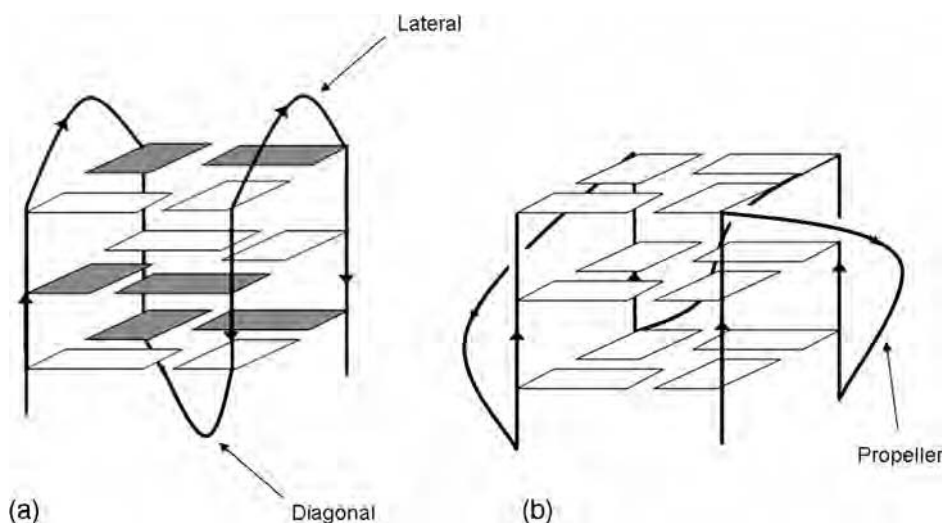


Figure 4.22 Schematic views of the folding of DNA strands in the intramolecular G-quadruplex formed from the human telomeric sequence $d[AGGG(TTAGGG)_3]$ with three G-tetrads: (a) the topology formed in sodium ion solution, as determined by NMR (Wang and Patel, 1993), and (b) the topology as found in the crystal structure obtained from potassium-containing solution (Parkinson, Lee, and Neidle, 2002).

The fundamental building block of all quadruplex structures is the G-tetrad. They are stacked one on another in a quadruplex structure, with a minimum of two being required for structural stability. They tend to have a (quasi) helical twist of ca 36°, giving the exterior of many quadruplexes an appearance superficially similar to a stretch of B-DNA double helix. The number of guanines in each individual G-tract is often (though not invariably) constant, and thus can be related directly to the number of G-tetrads formed in the folded quadruplex. Thus, in vertebrate telomeric DNA, where the repeat sequence is d(TTAGGG), the quadruplexes formed by either two or four repeats have three stacked G-tetrads.

The four guanosine nucleosides in an individual tetrad can in principle exist in either *anti* or *syn* glycosidic angle conformations, and thus there are 16 possible combinations (Fig. 4.23). The mutual orientation of individual strands in a quadruplex has consequences for the glycosidic angles. Thus, an all-parallel orientation for all four strands, as in the tetramolecular quadruplex shown schematically in Fig. 4.21, requires all glycosidic angles to adopt an *anti* conformation. Antiparallel quadruplexes can have a variety of orientations for the four backbones, whether these are from four separate strands or from a single (i.e. intramolecular) folded strand. These have both *syn* and *anti* guanines, arranged in a way that is particular for a given topology and set of strand orientations (Figs. 4.22a,b and 4.24).

The counter-ions play a key role in maintaining the integrity of quadruplex structures. They do this by forming ionic interactions with the O6 atoms of the tetrad guanines. At one extreme an ion can be positioned symmetrically in the plane of four guanines in a tetrad. Potassium ions tend to be situated symmetrically between two consecutive tetrads so that they are eight coordinated, to all eight O6 atoms from both tetrads (Fig. 4.25). This arrangement is favored by potassium and ammonium ions. The smaller sodium ion is able to coordinate all four O6 atoms in a single tetrad while being in the plane of all four guanines, although it can also be accommodated midway between tetrads. A continuum between both extreme positions can be seen in several crystal and NMR structures (e.g., Schultze *et al.*, 1999; Horvath and Schultz, 2001), with the formation of a continuous channel of positively charged counter-ions along the inner axis of the quadruplex (Fig. 4.26).

The variable sequences between the individual short G-tracts in a quadruplex generally form extrahelical loops. Adjacent linked parallel strands require a connecting loop to link the bottom G-tetrad with the top G-tetrad, leading to propeller-type loops (sometimes termed strand-reversal loops) – see Fig. 4.22b. Quadruplexes are designated

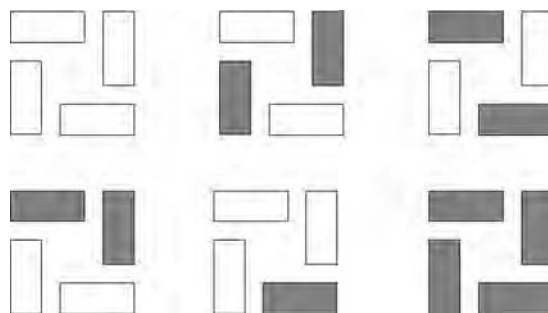


Figure 4.23 Schematic showing 6 of the 16 possible arrangements for the glycosidic angles of individual guanines in a G-tetrad. Shaded rectangles represent *syn* conformations, and unshaded rectangles represent *anti* arrangements. This convention is used in several of the figures in this section.

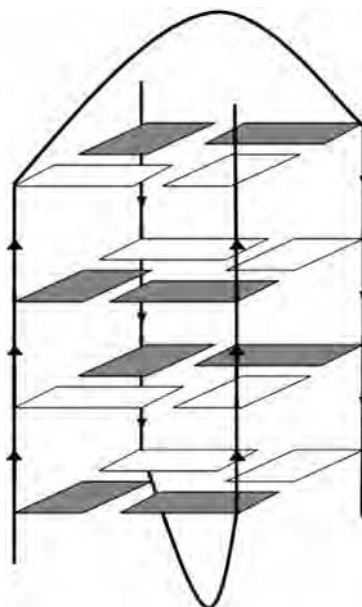


Figure 4.24 Schematic of the topology for the dimeric quadruplex formed by two strands of d(GGGGTTTGGGG), as found in the crystal structure (Haider, Neidle, and Parkinson, 2003). The rectangles, which represent guanosine nucleosides, are shaded as detailed in Fig. 4.23.

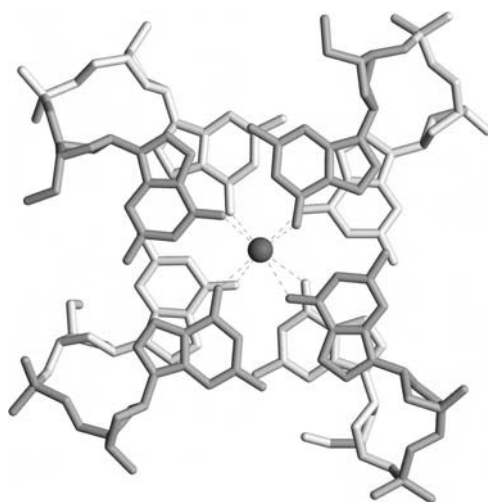


Figure 4.25 The octahedral coordination of a potassium ion, symmetrically positioned between two G-tetrad planes. The dashed lines indicate ionic interactions with O6 atoms of the guanines.

as antiparallel when at least one of the four strands is antiparallel to the others. This category of topology is found in the majority of bimolecular and in a number of intramolecular quadruplex structures determined to date. Two further types of loops have been observed in these structures, in addition to parallel loops. Lateral (sometimes termed edgewise) loops join adjacent G-strands (Fig. 4.22a). Two of these loops can be located

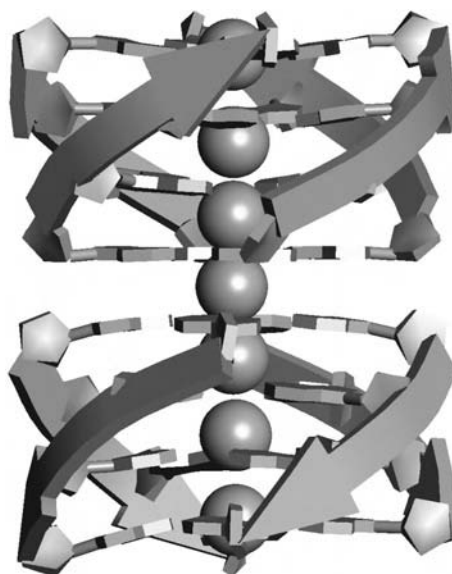


Figure 4.26 The channel of sodium ions (shown as spheres), running through the center of the dimer of two parallel d(TGGGGT) quadruplexes, as found in the crystal structure (Phillips *et al.*, 1997).

either on the same or opposite faces of a quadruplex, corresponding to head to head or head to tail respectively when in bimolecular quadruplexes. The second type of antiparallel loop, the diagonal loop joins opposite G-strands (Fig. 4.22a). In this instance the directionalities of adjacent strands must alternate between parallel and antiparallel, and are arranged around a core of four stacked G-tetrads.

All quadruplex structures have four grooves, in contrast with the two of a double helix. These are defined by the cavities bounded by the phosphodiester backbones. Groove dimensions are variable, and depend on overall topology and the nature of the loops, as well as glycosidic angles. Grooves in quadruplexes with only lateral or diagonal loops are structurally simple, and the walls of these grooves are bounded by monotonic sugar-phosphodiester groups. By contrast, grooves that incorporate propeller loops have more complex structural features that reflect the insertion of the variable-sequence loops into the grooves.

There are therefore a number of structural variables (number of G-tetrads; loop type, sequence and length; strand polarity), that together result in quadruplexes having high topological and structural variety. As yet, few general rules have emerged to give confidence in structure prediction, not least because there are insufficient experimental structures (Table 4.5) available for this to be reliable. It is apparent though that loop size and sequence is a major determinant. In particular, short sequences with ≤ 2 nucleotides are more likely to form propeller loops (for purely steric reasons), and thus impose a parallel-strand topology on that part of a quadruplex (Hazel *et al.*, 2004). Longer sequences are more likely to prefer lateral or diagonal loops. In these instances groove width may become enlarged as a consequence of the presence of the bulky loops. Even apparently minor changes in loop size (or in flanking 5' or 3' sequences), can have dramatic consequences for overall topology, as is illustrated below.

Table 4.5 Crystal and NMR Structures of Nucleic Acid Quadruplexes

Sequence	PDB code No.	NDB code No.
d(GGGGTTTTGGGG) in K ⁺	1JPQ	UD0013
d(GGGGTTTTGGGG) in Na ⁺ (X-ray)	1JB7	PD0218
d(GGGGTTTTGGGG) in Na ⁺ (NMR)	156D	-
d(TGGGGT)	352D	UDF062
d(GCATGC)	184D	UDG028
d(AGGGTTAGGGTTAGGGTTAGGG) in K ⁺ (X-ray)	1KF1	UD0017
d(AGGGTTAGGGTTAGGGTTAGGG) in Na ⁺ (NMR)	143D	-
d(UAGGGUTAGGGT)	1KBP	UD0016
d(GGGGTTTTGGGG)	2AVH	UD0062
d(GGGTTTTGGGG)	1U64	-
d(GGGGTTTTGGG)	1LVS	-
d(TGAGGGTGGIGAGGGTGGGGAAGG)	2A5P	-
d(GGGGTGGGAGGAGGGT)	1YBD	-
d(TTGGGTAGGGTTAGGGTTAGGGA)	2GKU	-
d(AAAGGGTAGGGTTAGGGTTAGGGAA)	2HY9	-
d(GTTAGGGTTAGGGT) + d(TAGGGU)	2AQY	-
d(GGGCGCGGGAGGAATTGGGCGGG)	2F8U	-
d(AGGGAGGGCGCTGGGAGGAGGG)	203M	-

4.3.3 Examples of Simple Quadruplex Structures

A number of sequences have been studied by NMR methods; some, such as d(TGGGGT) (Aboul-ela, Murchie, and Lilley, 1992) and d(TTGGGG) (Wang and Patel, 1992), form parallel-stranded quadruplex structures with all-*anti* glycosidic angles (Fig. 4.21). Others such as d(GGTTTTTCGG) (Wang *et al.*, 1991), form antiparallel arrangements with alternating *syn-anti* glycosidic bonds. The parallel d(TGGGGT) quadruplex structure, stabilized by sodium ions, has also been studied in detail by X-ray crystallography (Phillips *et al.*, 1997). The structure, which contains four independent quadruplexes in the crystallographic asymmetric unit, has been solved to exceptionally high resolution (0.95 Å), enabling many details of the structure to be precisely defined. Two quadruplexes are stacked together in the crystal, with a channel of seven sodium ions running through the center of the structure (Fig. 4.26). These occupy a range of positions for these ions (and hence of ion-ion distances) with respect to the quartet planes, although there is a clear pattern of increasing coplanarity with the quartets toward the ends of the tetraplex dimer. Most of the sugar rings are in the normal B-DNA C3'-*endo* conformation. However the distribution of values for the backbone torsion angles δ and ϵ is outside the normal duplex ranges, suggesting that the quadruplex backbone is strained, possibly in order for the hydrogen-bonding of the tetrads to occur.

The sequence d(GGGGTTTTGGGG), from the telomere of the *Oxytricha nova* organism, has been extensively studied by structural methods. The general arrangement is of a four-stranded, folded quadruplex with two strands associating together as an intramolecular hairpin structure. There are four stacked hydrogen-bonded G-quartets (Figs. 4.24, 4.27), sandwiched on top of each other at a separation of 3.4 Å. Several crystallographic analyses have been reported for this sequence; with sodium ions and cocrystallized within the structure of a protein-single-stranded telomere complex (Horvath and Schultz, 2001), and crystallized as the native structure in the presence of potassium ions (Haider, Parkinson, and Neidle, 2002). Both structures show the quadruplex with adjacent strands antiparallel to each other with the loops joining diagonal strands. There are alternating *syn* and *anti* glycosidic conformations along



Figure 4.27 Ribbon representation of the crystal structure of the dimeric quadruplex formed by two strands of d(GGGGTTTTGGGG) (Haider, Neidle, and Parkinson, 2003).

any one strand (Fig. 4.24), and the arrangement of strands results in a G-quartet arrangement with a *syn-syn-anti-anti* pattern of glycosidic angles and loops joining diagonal strands. This structure is in agreement with a series of NMR studies (Smith and Feigon, 1992; Schultze *et al.*, 1999) on this sequence with sodium, potassium, and ammonium ions. The overall fold is thus retained in all of these environments. There are some differences between the NMR and crystal structures that are mostly concerned with the conformations of thymines in the loops. In the latter the thymines are highly ordered and there is extensive stacking between them (Fig. 4.27).

Apparently minor changes to the *Oxytricha* sequence d(GGGGTTTTGGGG) result in quadruplexes with major structural changes. Two distinct bimolecular quadruplexes are formed in the crystal by d(G₄T₃G₄), with one being a head-to-tail lateral-loop dimer in which all adjacent strands are antiparallel, and the other is a head-to-head hairpin quadruplex with one adjacent strand parallel and the other antiparallel (Hazel, Parkinson, and Neidle, 2006). Decreasing the size of one G-tract by one guanosine at the 5'-end, so that the sequence becomes d(GGGTTTTGGGG), results in a bimolecular quadruplex having both lateral and diagonal loops (Šket, Črnugelj, and Plavec, 2003). This is one of the few cases where a bimolecular quadruplex has an unequal number of parallel (three) and antiparallel (one) strands (Fig. 4.28) (Table 4.5).

4.3.4 Some Complex Quadruplex Structures

A complex pattern of folding has been found in the intramolecular foldback quadruplexes (Figs. 4.22 and 4.29) formed by the sequence d[AGGG(TTAGG)₃], consisting of almost four repeats of the vertebrate telomeric sequence d(TTAGG). The NMR structure in sodium solution (Wang and Patel, 1993) shows each strand flanked by one parallel and one antiparallel neighbors. There is a core of three guanine

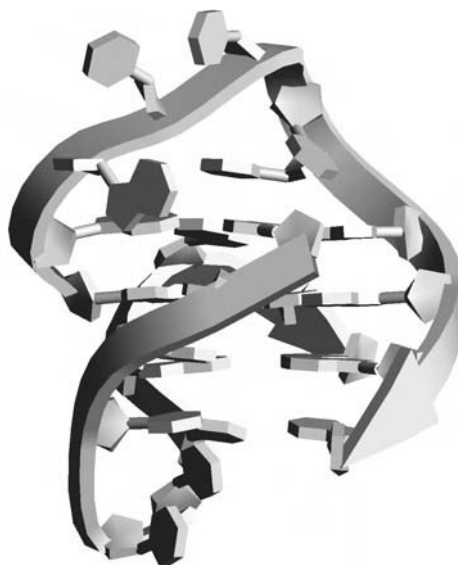


Figure 4.28 Ribbon representation of the crystal structure of the asymmetric dimeric quadruplex formed by two strands of d(GGGTTTGGGG) (Šket, Črnugelj, and Plavec, 2003).

tetrads flanked by two lateral and one diagonal TTA loop. The asymmetry of the loops results in significant differences between the width of the four grooves (Fig. 4.29). For the NMR structure with the lowest rmsd, two grooves have intermediate widths of 11.5–13.5 Å, and are flanked by a narrow (<10 Å) and a wide groove (>15 Å). It has been suggested that these differences may be of significance for the recognition of quadruplexes by telomere-binding proteins (Arthanari and Bolton, 2001).

The crystal structure of the same sequence, grown from potassium-containing solution, shows a remarkably different topology (Figs. 4.22b and 4.30), with all strands parallel and with all three TTA loops arranged in a propeller-like arrangement (Parkinson, Lee, and Neidle, 2002). The conditions used to form these crystals have some

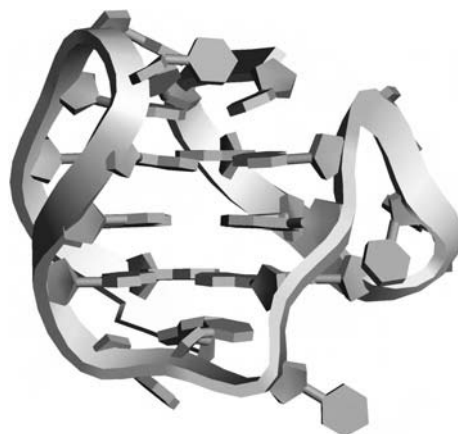


Figure 4.29 Representation of the NMR structure (Wang and Patel, 1993) of the sodium form of the intramolecular quadruplex formed by d[GGG(TTAGG)₃].

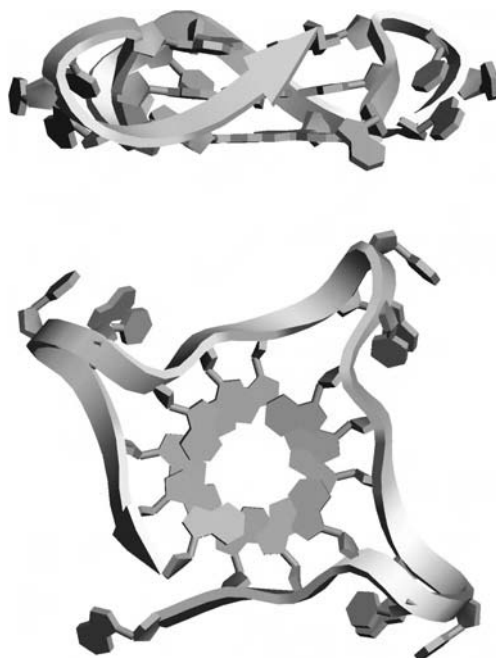


Figure 4.30 Two views of the crystal structure (Parkinson, Lee, and Neidle, 2002) of the potassium form of the intramolecular quadruplex formed by d[GGG(TTAGGG)₃]. The top view shows the narrow cross section of this G-quadruplex topology; the bottom view shows the three propeller loops and their closely similar conformations.

similarities to those in the cell nucleus, with a typical potassium ion concentration of ca 100 mM, so it has been suggested that this topology is relevant to the folded state of the 3′-single-stranded telomeric DNA overhang, in apparent accord with circular dichroism studies (Rujan, Meleney, and Bolton, 2005). However, it is also apparent from a number of biophysical studies (e.g., Ying *et al.*, 2003) that the 22-mer in potassium solution can exist as a mixture of several conformers; the topology found in the crystal structure is one low-energy form that is favored by G-quadruplex/G-tetrad stacking, and perhaps by the sheer density of a crystal structure. Other topologies have been observed in NMR studies of several sequences related to the 22-mer, with short 3′ and/or 5′ flanking sequences that serve to stabilize particular structures, especially one with (3+1) loops (Luu *et al.*, 2006; Ambrus *et al.*, 2006; Dai *et al.*, 2007). The (3+1) fold is also an unusual and unexpected topology, with one antiparallel and three parallel strands, so that there are one propeller and two lateral loops with two *syn-anti-anti-anti* and one *anti-syn-syn-syn* G-tetrads (Fig. 4.31). What remains still to be determined by fine-structure methods is the precise nature of *all* the species present in the K⁺-solution of d[AGGG(TTAGGG)₃] and related sequences, and then how these relate to the structure of the single-stranded telomeric DNA overhang, whose length suggests that it is capable of folding into several consecutive quadruplex structures.

The intramolecular quadruplex structures available to the G-rich strand of nuclease-hypersensitive NHE III1 promoter sequence in the *c-myc* oncogene have been characterized by NMR methods. There are more than four short G-tracts overall in this sequence, and structural studies have been undertaken on quadruplexes with either four or five such tracts. The overall folds are closely similar, with all-parallel topologies and therefore all loops constrained to be in propeller conformations (Phan, Modi, and



Figure 4.31 A view of the (3+1) intramolecular quadruplex formed by the sequence d(TTGGGTTAGGGTTAGGGTTAGGGA), as determined in potassium solution by NMR methods, showing the one propeller and two lateral loops (Luu *et al.*, 2006)

Patel, 2004; Ambrus *et al.*, 2005; Phan *et al.*, 2005). A recurrent theme in these structures is the existence of a platform that stacks onto a tetrad end, providing additional stabilization (Fig. 4.32). Related platforms have subsequently been found in several of the (3+1) telomeric quadruplexes, where they are formed by the flanking sequences, and perform the same stabilizing function. A more unusual quadruplex fold has been found (Phan *et al.*, 2007) in a sequence that occurs in the *c-kit* kinase promoter region (Rankin *et al.*, 2005). This structure (Fig. 4.33) has an isolated guanine, previously assumed to be part of a loop tract, being involved in G-tetrad core formation, despite the presence of four three-guanine tracts. There are four loops: two single-residue propeller loops, a two-residue loop, and a five-residue stem-loop at one end of the core quadruplex, which contain two A•G base pairs.

The realization that quadruplex structures could be formed by bases other than solely guanine has led to the determination of several remarkable folded DNA structures



Figure 4.32 One of the *c-myc* NMR-determined quadruplex structures, for the sequence d(TGAGGGTGGIGAGGGTGGGGAAGG) (Phan *et al.*, 2005). The nucleotide platforms are apparent above and below the three G-tetrads that constitute the quadruplex core.



Figure 4.33 NMR structure of the quadruplex formed by the G-rich sequence d(AGGGAGGGCGCT GGGAGGAGGG), which occurs in the promoter sequence of the human c-kit kinase (Phan *et al.*, 2007).

(Patel *et al.*, 1999). Such structures may be formed by G-rich sequences when used as aptamers for highly specific binding to protein targets. The short sequence d(GCATGC) forms a loop arrangement which associates into an antiparallel dimer (Fig. 4.34) with two CGCG quartets (Leonard *et al.*, 1995). Adenines have been found (Kettani *et al.*, 2000a) to associate with a conventional G-quartet in the structure of the sequence d(GGAGGAG) to form A•G•G•G•G•A hydrogen-bonded hexads stacked between two G-quartets (Fig. 4.35). Although many more possible arrangements for these types of mixed G-rich sequences remain to be thoroughly explored, it is now possible to design appropriate sequences that will incorporate features such as quartets and triplets in defined relationships to each other (Kettani *et al.*, 2000b; Kuryavyi *et al.*, 2001).

Molecular dynamics simulations of G-quadruplexes over multi-nanosecond timescales have shown that the G-quartet arrangement is remarkably stable, especially in the parallel d(TGGGGT) quadruplex (Spáčková, Berger, and Šponer, 1999; Read and Neidle, 2000). Simulations have included sodium and potassium counter-ions as well as water molecules, and so have been able to ascertain the roles played by ions in both conventional hairpin dimers with standard G-quartets (Strahan, Keniry, and Shafer, 1998), and in quadruplexes with mixed GCGC quartets (Spáčková, Berger, and Šponer, 2001). The latter appear to require just two cations for a structure comprising

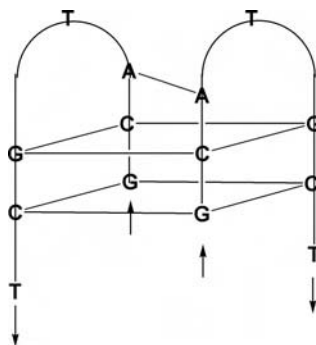


Figure 4.34 Schematic representation of the quadruplex formed by the sequence d(GCATGCT), as found in its crystal structure (Leonard *et al.*, 1995).

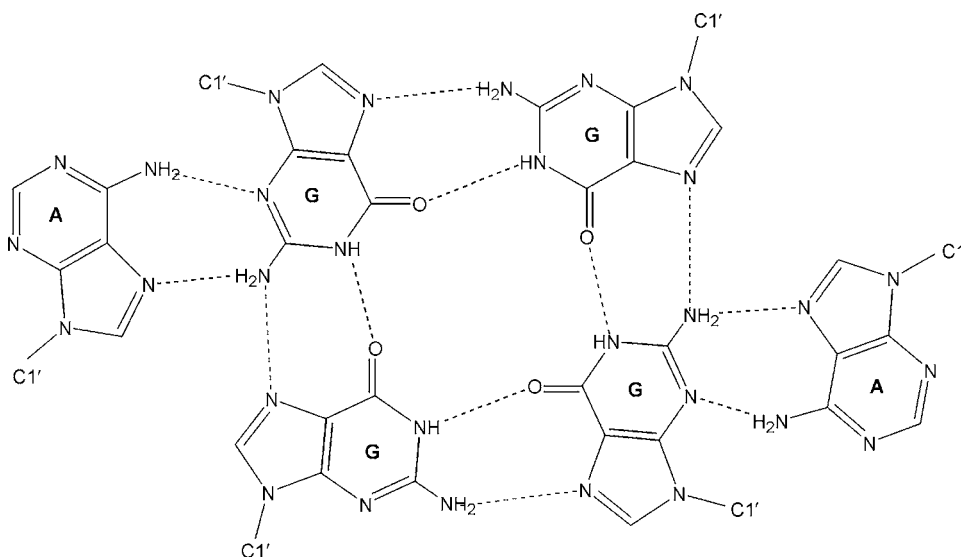


Figure 4.35 The hydrogen-bonding postulated for the A•G•G•G•G•A hexad (Kettani *et al.*, 2000).

two G-quartets and two CGCG ones, with subtle differences in the hydrogen-bonding arrangement being dependent on the nature of the cation.

Some general patterns of the relationships between strand polarity and base-pairing, have emerged from studies on triplexes and quadruplexes, as well as on Watson–Crick, left-handed and mispaired duplexes (Westhof, 1992; Lavery *et al.*, 1992). For example, the pattern of glycosidic angles for a base pair is a consequence of the relationship between strand directions in quadruplexes, with *syn* angles being produced by anti-parallel strands and *anti* angles by parallel ones. This has led to a general classification scheme for nucleic acid structural types (Lavery *et al.*, 1992), based on these polarity and base-pairing criteria. Prediction of new structural types emerges naturally from the scheme. The variety of DNA structures that have been established over the past few years and described in this Chapter suggests that many more remain to be experimentally determined, especially for quadruplex nucleic acids.

4.3.5 The i-Motif

Diffraction studies on poly(C) fibers produced at acidic pH were interpreted (Langridge and Rich, 1963) in terms of a parallel-stranded duplex with C•C⁺ mismatch base pairs (Fig. 4.36). As with guanine-quadruplexes, this type of arrangement was largely

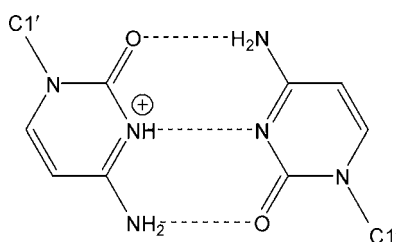


Figure 4.36 The hydrogen-bonding arrangement in a C•C⁺ base pair.

Table 4.6 Crystal and NMR Structures of Nucleic Acid I-Motifs

Sequence	PDB code no.	NDB code no.
d(CCCC)	190D	UDD024
d(CCCT)	191D	UDD023
d(ACCCT)	1BQJ	UD0003
d(TAACCC)	200D	UDF027
d(CCCAAT)	241D	UDF043
d(AACCCC)	294D	UDF054
d(CCTTTCCCTTTACCTTTCC)	1A83	-
d(CCCTAA ^{5me} CCCTAACCCUAACCCT)	1EL2	-

forgotten for some years. An NMR study on the cytosine-rich oligonucleotide d(TCCCC) showed that this sequence formed, not a duplex, but instead a novel four-stranded quadruplex (Gehring, Leroy, and Guéron, 1993), which was termed the i-motif. The basic unit is the C•C⁺ mismatch base pair, with one of the two cytosine protonated at the N3 position such that three hydrogen bonds are formed (Table 4.6).

The arrangement of the i-motif bears some resemblance to the original fiber diffraction model in that the two strands connected by the C•C⁺ base pairs are in a parallel orientation. Two of these duplexes are intercalated into each other that results in sets of consecutive C•C⁺ base pairs oriented approximately at right angles to each other (Fig. 4.37). The DNA strand complementary to a guanine-rich double-stranded centromeric or telomeric sequence is by definition cytosine-rich. A number of crystal structures of such sequences have been reported, all of which show this structural motif (Kang *et al.*, 1994, 1995; Berger *et al.*, 1995; Cai *et al.*, 1998). These show that the grooves in the i-motif structures are highly asymmetric, with two broad, shallow major grooves, of width typically 16 Å. Each is flanked by a very narrow minor groove (Fig. 4.38), of width around 6 Å. The helical twist between consecutive C•C⁺ base pairs in either duplex tends to be small, between 12° and 20°.

The i-motif sequence motif occurs naturally in duplex telomeric DNA, as the complement to the quadruplex-forming G-rich strand. The NMR structure has been determined (Phan, Guéron, and Leroy, 2000), for the human telomeric i-motif sequence d(CCCTAA^{5me}CCCTAACCCUAACCCT). This forms an intramolecular i-motif structure with three lateral loops joining the strands (Fig. 4.39), so that there are two narrow and two wide grooves.

4.4 DNA Junctions

4.4.1 Holliday Junction Structures

The process of homologous genetic recombination involves the crossing-over of strands from two DNA sequences. The central intermediate in this process is believed to be a four-way branched junction-type structure (Fig. 4.40), termed a Holliday junction. It consists of four strands, each participating in four antiparallel base-paired arms. A wide variety of biophysical techniques, such as gel electrophoresis and fluorescence energy transfer have been used in attempts to define the three-dimensional structure of these junctions (Lilley and Norman, 1999). Data from these studies (Duckett *et al.*, 1995), together with molecular modeling (Von Kitzing, Lilley, and Diekmann, 1990),

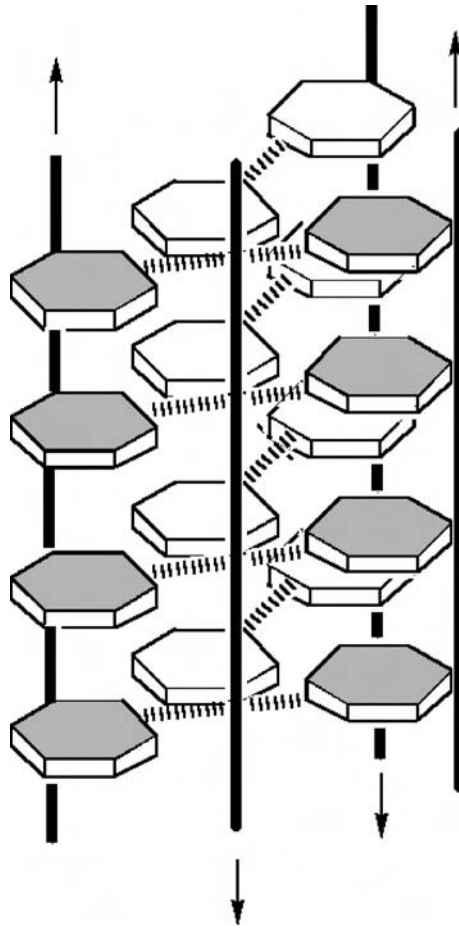


Figure 4.37 Schematic view of the arrangement of base pairs in an i-motif four-stranded structure. The base pairs shown shaded form one double helix, which is intercalated into that formed by the unshaded base pairs.

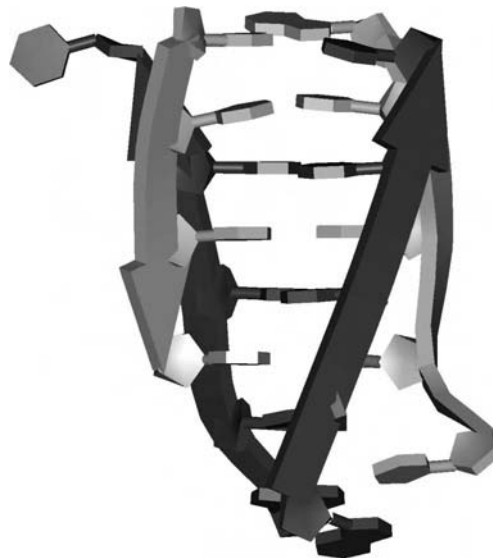


Figure 4.38 Structure of an i-motif sequence d(ACCCT), as found in its crystal structure (Weil *et al.*, 1999).

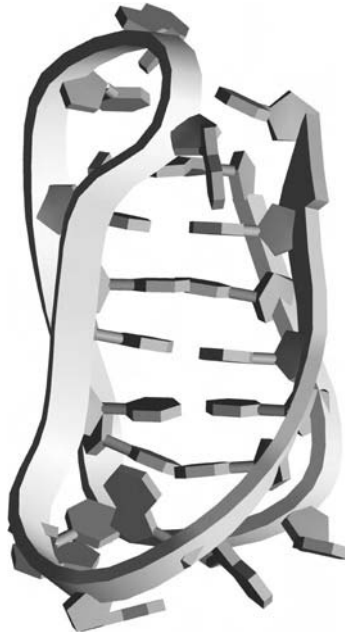


Figure 4.39 The NMR structure (Phan, Guéron, and Leroy, 2000), for the human telomeric i-motif sequence d(CCCTAA^{5me}CCCTAACCCUAACCCT).

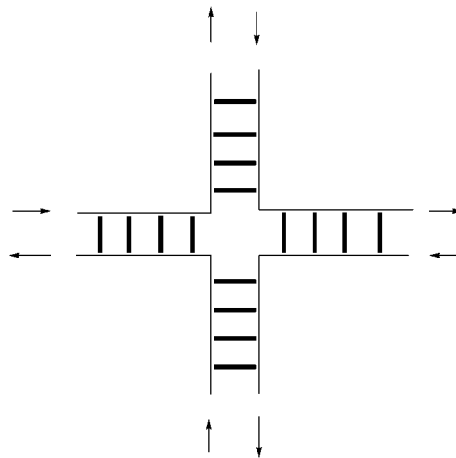


Figure 4.40 The organization of a four-way Holliday junction. The directionality of each of the four strands is shown by arrows.

resulted in the proposal of the X-structure, with two pairs of almost parallel helices and a region in the center where the DNA strands from the two starting duplexes cross over.

There have been numerous attempts to crystallize junction-type DNA structures (Lilley and Norman, 1999), mostly involving long sequences. Success was finally achieved independently in two laboratories, very surprisingly with two short, closely similar decamer sequences (Table 4.7). The first, d(CCGGGACCGG), was crystallized

Table 4.7 Crystal Structures of Nucleic Acid Junctions

Sequence	PDB code No.	NDB code No.
d(CCGGTACCGG)	1DCW	UD0008
d(CCGGGACCGG)	467D	UD0006
d(TCGGTACCGA)	1NQS	UD0023
10-23 DNA enzyme + RNA	1BR3	UH0001
10-23 DNA enzyme + RNA	1EGK	UH0002
d(GGAC ^{Bz} AGAU ^{Bz} GGGAG)	1P1Y	UD0031

(Ortiz-Lombardia *et al.*, 1999) with the goal of studying contiguous G•A mismatches in duplex DNA. The crystal structures of the fully inverted repeat sequences d(CCGGTACCGG) and d(CCGCTAGCGG) shows a closely similar arrangement (Eichman *et al.*, 2000), albeit without any possible perturbing effects of the G•A mismatch. All three structures consist of two pairs of arms of the junction stacked on each other to form two continuous antiparallel duplexes that have totally normal linear B-DNA conformations (Figs. 4.41 and 4.42). All bases are in base-paired arrangements in the duplex regions. The junction itself is formed by a few changes in backbone conformations, chiefly in the two cytosine-7 nucleotides on the two inner-facing strands. The two arms are right-hand twisted with an angle between them of $\sim 40^\circ$, in accord with the stacked X-junction model – it is remarkable how close the crystal structure matches the conclusions from the earlier biophysical and modelling studies. The 5'-ACC sequence is at the core of the junction and may be important for its stabilization. The closeness of four phosphate groups at the junction requires compensating cationic charge, and a sodium ion has been located in a buried position within the junction, bridging two adenosine-6 phosphate oxygen atoms and situated on the local twofold axis of the structure. Crystal structures of the Holliday junction structure formed by the sequence d(TCGGTACCGA) with sodium, calcium, and strontium ions (Thorpe *et al.*, 2003), have provided detailed views of the roles played by ions, not only at the junction itself, but also in the helical stems.

Comparative analysis of several junction structures (Eichman *et al.*, 2002) has shown that forming the junction has little impact on the sequence-dependent B-form structure of the duplex DNA arms, or the pattern of hydration in the duplex grooves. A general methodology for analyzing and comparing the geometry of Holliday junction structures

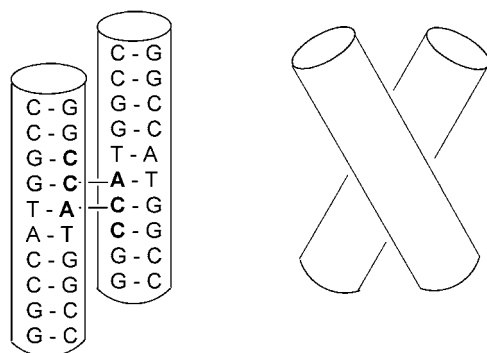


Figure 4.41 Schematic views of a Holliday junction structure, showing, on the left, the two helical arms and their connection points via the two central crossover sequences. The relative angle of the two helical arms is shown on the right.

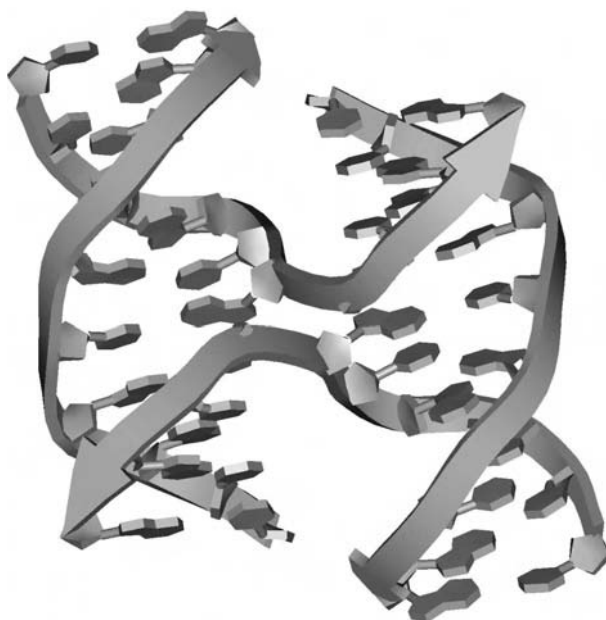


Figure 4.42 Crystal structure of the Holliday junction formed by four decamer strands (Eichman *et al.*, 2000).

has been proposed, that provides descriptors for the relative orientation, rotation, and translation of the duplex helical arms (Watson, Hays, and Ho, 2004).

The crystal structure (Paukstelis *et al.*, 2004) of the 13-mer sequence d(GGAC^{Br}AGAU^{Br}GGGAG) shows that each individual monomer strand base pairs with three crystallographically-related strands. Part of the monomer is involved in a short antiparallel double helix (Fig. 4.43). The monomer then has a sharp turn in its backbone conformation which enables the other part of the strand to form a junction of parallel-stranded homopurine base pairs with a strand from an adjacent layer. The net effect is to form a unique continuous three-dimensional lattice structure with large solvent channels. This structure has been used as the basis for the design of DNA nanostructures such as molecular sieves (Paukstelis, 2006).

4.4.2 DNA Enzyme Structures

More complex junction structures have been found in the crystal structures of “DNA enzymes.” These analyses follow the discovery, using *in vitro* selection procedures, that certain DNA sequences can act as enzymes, for example being able to cleave phosphodiester bonds (Breaker and Joyce, 1994). The “10–23” motif, which catalyzes the cleavage of RNA sequences, has been cocrystallized (Nowakowski *et al.*, 1999a) with an inactivated RNA substrate. The resulting crystal structure consists of a dimer of the DNA:RNA complex in which there are two DNA and two RNA strands. These form five double-helical regions (Figs. 4.44). A four-way junction is produced in this structure by the confluence of the hexanucleotide loop with three stem regions. As in



Figure 4.43 Crystal structure of two strands of the 13-mer sequence d(GGAC^{Br}AGAU^{Br}GGGAG) (Paukstelis *et al.*, 2004), showing the sharp turn approximately half-way along the backbone that results in distinct types of helix being formed in the crystal lattice.

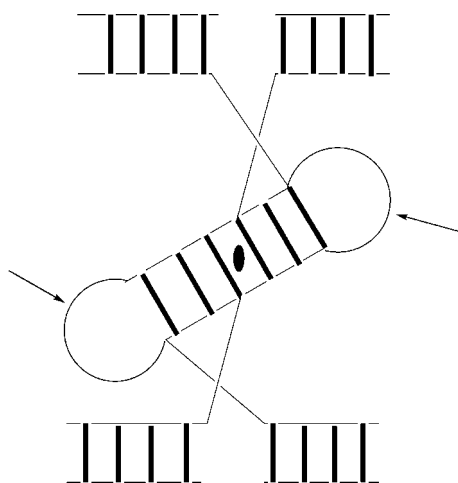


Figure 4.44 Schematic view of the connectivity of the helices and loop regions in found in the crystal structure of the 10–23 DNA enzyme (Nowakowski *et al.*, 1999).

the Holliday junction structures, all bases are base-paired and stacked. All of these junction structures have an overall X-form, and there are few differences between them, apart from the A-form nature of two stems in the 10–23 junction (which arises from the involvement of RNA). By contrast, a 108-nucleotide DNA-RNA complex (Nowakowski *et al.*, 1999b) adopts an X-junction structure but with a left-handed orientation of the arms (Fig. 4.45).

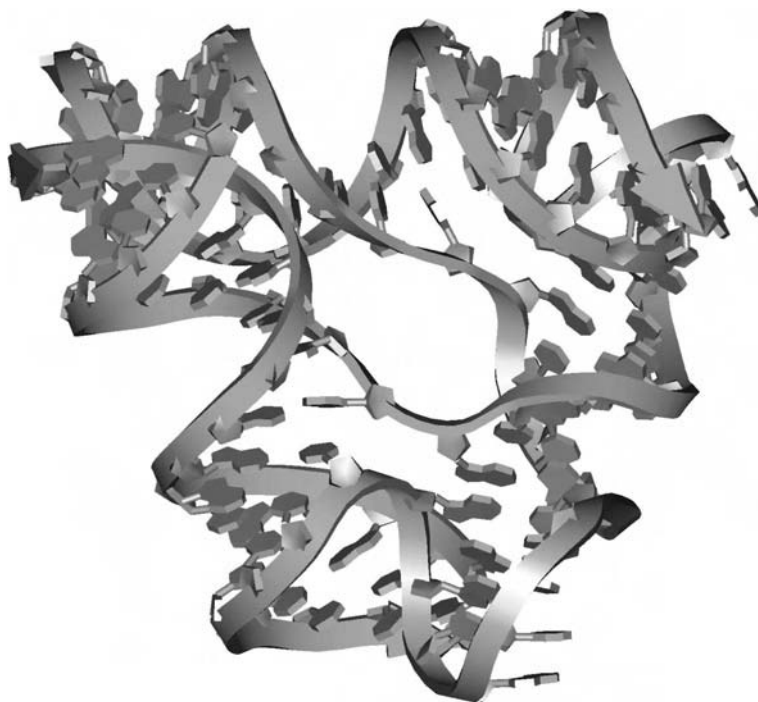


Figure 4.45 Crystal structure of the 108-mer DNA enzyme (Nowakowski *et al.*, 1999).

4.5 Unnatural DNA Structures

A very large number of chemically modified DNAs have been reported in the literature, many of which have been constructed for use as antisense probes. There is a significant crystallographic literature on some of these DNA analogues (see Egli and Pallan, 2007 for a comprehensive survey). Modifications that have been studied cover all possible functional groupings in an oligonucleotide, ranging from backbone changes through to deoxyribose and base alterations. We focus here on some of the more structurally studied ones, of which the peptide nucleic acid (PNA) backbone is of particular importance as a biological probe (Nielsen *et al.*, 1991; Nielsen, 2004; see also Sect. 4.2.2).

Two crystal structures of PNAs within triplex structures have been reported (Sect. 4.2.2), as well as a small number of crystal and NMR structures of PNA-containing duplexes. One such crystal structure (Rasmussen *et al.*, 1997) is of a hexamer duplex with the base sequence CGTACG and with pure PNA backbones in both strands. The helix (Fig. 4.46) is exceptionally open, with a 28 Å width and an 18 Å pitch; although the base pairs are perpendicular to the helix axis, this is far from a classic B-form. A more complex decamer duplex PNA-DNA structure (Menchise *et al.*, 2003), has one conventional DNA strand and a complementary antiparallel strand incorporating three chiral D-lysine monomers (Fig. 4.47). This helix also has a wide, deep major groove and a correspondingly shallow and narrow minor groove, with the base pairs perpendicular to the helix axis. The helix type adopted by PNAs is conformationally more restricted than with standard backbones, and has been termed the P-form, by contrast with the standard A- and B-forms. The complexity of PNA structures has also

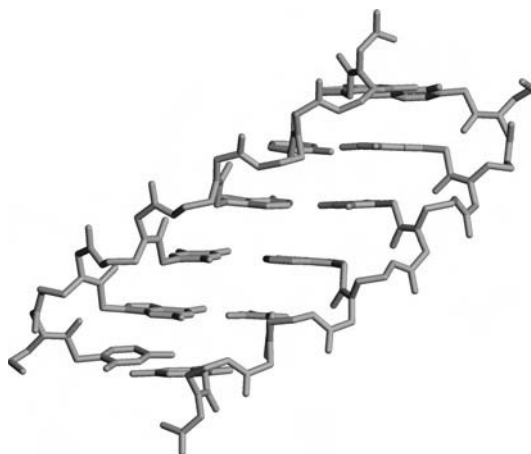


Figure 4.46 Crystal structure of a PNA duplex (Rasmussen *et al.*, 1997).



Figure 4.47 Crystal structure of a decamer duplex PNA-DNA (Menchise *et al.*, 2003). The backbone of the DNA strand is shown in ribbon representation, and that of the PNA in stick form.

been illustrated by the formation of both left- and right-handed helices in the same PNA crystal structure (Petersson *et al.*, 2005).

Replacement of natural DNA bases with synthetic analogues has long been an active area of study. Expanded-size bases have been explored as the basis for an alternative genetic system, which can be recognized by the standard replication apparatus of the cell. Base extension with the addition of a benzene ring (see Fig. 4.48) raises the ques-

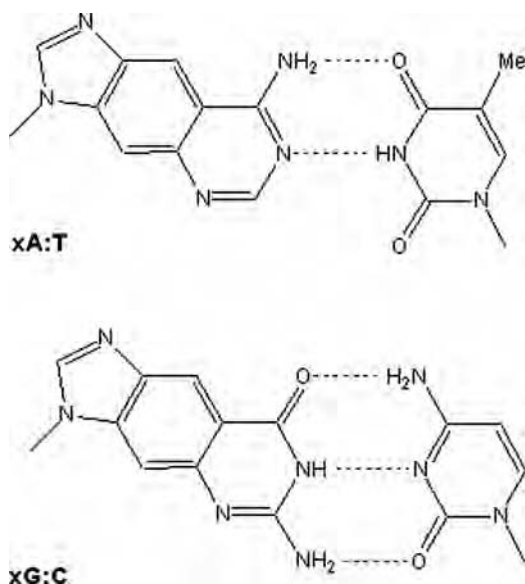


Figure 4.48 xA•T and xG•C base pairs, showing the purines extended with a benzene ring to form xDNA bases.

tion of whether a DNA constructed from such “x” base pairs can still form a conventional double helix, a prerequisite for biological functioning. The NMR structure of an eight-base xDNA double helix (Lynch *et al.*, 2006) has been described. This sequence (Table 4.8) has been designed so that as many as possible nearest-neighbor effects for the xDNA bases are sampled. The resulting structure (Fig. 4.49) is a right-handed double

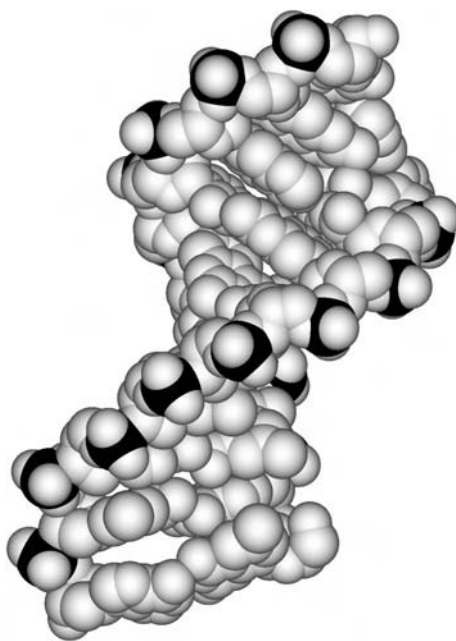


Figure 4.49 Space-filling representation of the NMR-derived structure of xDNA (Lynch *et al.*, 2006).

Table 4.8 Crystal and NMR Structures of Unnatural Nucleic Acids

Sequence	PDB code no.	NDB code no.
$d(C_{pn}G_{pn}T_{pn}A_{pn}C_{pn}G_{pn})$	1PUP	UPNA56
$^1d(pAGTGA TCTAC) + d(G_{pn}T_{pn}A_{pn}GATCA_{pn}C_{pn}T_{pn})$	1NRB	UP0003
$^2d(TG_xTA_xC_xGC_xA_xGT) + d(A_xCA_xTGC_xGTC_xA_x)$	2ICZ	UD0023
$^3d(C_hG_hA_hT_hT_hC_hG)$	2H9S	UD0070
$^4d(GCGTA_lACGC)$	1I5W	AD0020

pn indicates a peptide nucleic acid (PNA) backbone¹, replacing a standard backbone.
x indicates a benzylated base², replacing a standard base.
h indicates a dideoxy-β-D-glucopyranose sugar, replacing a standard deoxypentose sugar³.
l. indicates a locked nucleic acid sugar⁴, replacing a standard deoxypentose sugar.

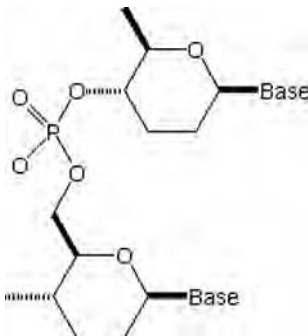


Figure 4.50 The dideoxyhexose arrangement in homo DNA, replacing the deoxyribose in conventional DNA.

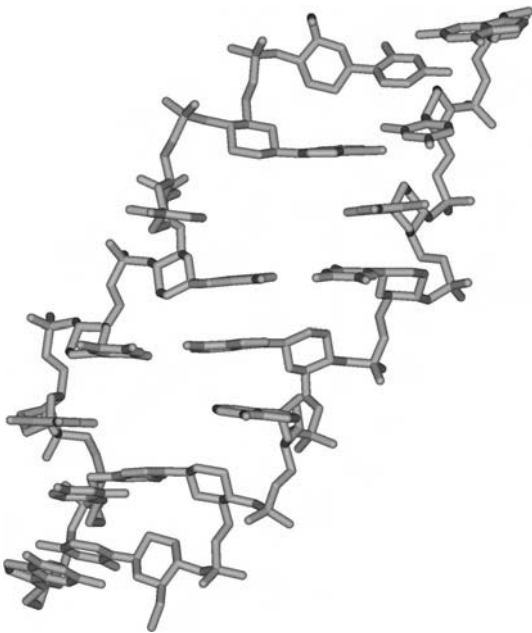


Figure 4.51 The crystal structure of the homo DNA duplex (Egli *et al.*, 2006). Note the irregular backbone and base morphology.

helix with all glycosidic angles in the *anti* range and all sugars having C2'-*endo* pucker. Although the helix has some resemblance to the B-form, there are significant differences in local and global morphology. The average helical twist is 30°, giving 12 residues per helical turn, and the average base-pair inclination to the helix axis is 24°. Consequently the minor groove is over 7 Å wider in xDNA than in B-DNA. Whether an xDNA can be replicated is as yet unknown. However the structural study has shown that even if DNA polymerase has difficulties in coping with the large size of xDNA bases, they are not an impediment to double helix formation (Table 4.8).

We conclude this chapter with a structure that addresses the question of why Nature has chosen five-membered pentose sugar rings as part of the building-blocks in nucleic acids rather than the more naturally abundant hexose motif. The crystal structure of a homo-DNA octamer duplex with all nucleosides having deoxyhexose rings replaced by deoxypentose (Fig. 4.50), has been determined (Egli *et al.*, 2006). The structure determination (itself a crystallographic tour de force) shows that although the sequence forms an antiparallel duplex helix (Fig. 4.51), the arrangement is highly irregular (and bears little resemblance to previous predictions of homo-DNA structure). Two bases are nonpaired and protrude out from the structure, and there is almost no base–base stacking within each strand. Base-step geometry is also highly irregular along the sequence. Overall the structure provides a rationalization of why homo-DNAs are unable to form stable pairing arrangements, and ultimately why Nature did not choose them.

References

- Aboul-ela, F., Murchie, A. I. H., and Lilley, D. M. J. (1992). NMR study of parallel-stranded tetraplex formation by the hexadeoxynucleotide d(TG₄T). *Nature*, **360**, 280–282.
- Ambrus, A., Chen, D., Dai, J., Bialis, T., Jones, R. A., and Yang, D. (2006). Human telomeric sequence forms a hybrid-type intramolecular G-quadruplex structure with mixed parallel/antiparallel strands in potassium solution. *Nucleic Acids Res.*, **34**, 2723–2735.
- Ambrus, A., Chen, D., Dai, J., Jones, R. A., and Yang, D. (2005). Solution structure of the biologically relevant G-quadruplex element in the human c-MYC promoter. Implications for G-quadruplex stabilization. *Biochemistry*, **44**, 2048–2058.
- Arnott, S., Bond, P. J., Selsing, E., and Smith, P. J. C. (1976). Models of triple-stranded polynucleotides with optimised stereochemistry. *Nucleic Acids Res.*, **10**, 2459–2470.
- Arnott, S., Chandrasekaran, R., and Marttila, C. M. (1974). Structures for polyinosinic acid and polyguanylic acid. *Biochem. J.*, **141**, 537–543.
- Arthanari, H. and Bolton, P. H. (2001). Functional and dysfunctional roles of quadruplex DNA in cells. *Chem. Biol.*, **8**, 221–230.
- Bates, P. J., Laughton, C. A., Jenkins, T. C., Capaldi, D. C., Roselt, P. D., Reese, C. B., and Neidle, S. (1996). Efficient triple helix formation by oligodeoxyribonucleotides containing α - or β -2-amino-5-(2-deoxy-D-ribofuranosyl) pyridine residues. *Nucleic Acids Res.*, **24**, 4176–4184.
- Beal, P. A. and Dervan, P. B. (1991). Second structural motif for recognition of DNA by oligonucleotide-directed triple-helix formation. *Science*, **251**, 1360–1363.
- Berger, I., Kang, C., Fredian, A., Lockshin, C., Ratliff, R., Moysis, R., and Rich, A. (1995). Extension of the four-stranded intercalated cytosine motif by adenine•adenine base pairing in the crystal structure of d(CCCAAT). *Nature Struct. Biol.*, **2**, 416–425.
- Betts, L., Josey, J. A., Veal, J. M., and Jordan, S. R. (1995). A nucleic acid triple helix formed by a peptide nucleic acid-DNA complex. *Science*, **270**, 1838–18341.
- Bijapur, J., Keppler, M. D., Bergqvist, S., Brown, T., and Fox, K. R. (1999). 5-(1-propargylamino)-2'-deoxyuridine (UP): a novel thymidine analogue for generating DNA triplexes with increased stability. *Nucleic Acids Res.*, **27**, 1802–1809.
- Bohon, J. and de los Santos, C. R. (2003). Structural effect of the anticancer agent 6-thioguanine on duplex DNA. *Nucleic Acids Res.*, **31**, 1331–1338.

- Bohon, J. and de los Santos, C. R. (2005). Effect of 6-thioguanine on the stability of duplex DNA. *Nucleic Acids Res.*, **33**, 2880–2886.
- Breaker, R. R. and Joyce, G. F. (1994). A DNA enzyme that cleaves RNA. *Chem. Biol.*, **1**, 223–229.
- Brown, T., Hunter, W. N., Kneale, G., and Kennard, O. (1986). Molecular Structure of the G–A base pair in DNA and its implications for the mechanism of transversion mutations. *Proc. Natl. Acad. Sci. USA*, **83**, 2402–2406.
- Cai, L., Chen, L., Raghavan, S., Ratliff, R., Moysis, R., and Rich, A. (1998). Intercalated cytosine motif and novel adenine clusters in the crystal structure of the *Tetrahymena* telomere. *Nucleic Acids Res.*, **26**, 4696–4705.
- Catapano, C. V., McGuffie, E. M., Pacheco, D., and Carbone, G. M. R. (2000). Inhibition of gene expression and cell proliferation by triple helix-forming oligonucleotides directed to the *c-myc* gene. *Biochemistry*, **39**, 5126–5138.
- Chandrasekaran, R., Giacometti, A., and Arnott, S. (2000a). Structure of poly (U).poly (A).poly (U). *J. Biomol. Struct. Dynamics*, **17**, 1023–1034.
- Chandrasekaran, R., Giacometti, A., and Arnott, S. (2000b). Structure of poly (dT).poly (dA).poly (dT). *J. Biomol. Struct. Dynamics*, **17**, 1011–1022.
- Chatake, T., Ono, A., Ueno, Y., Matsuda, A., and Takénaka, A. (1999a). Crystallographic studies on damaged DNAs. I. An N(6)-methoxyadenine residue forms a Watson–Crick pair with a cytosine residue in a B-DNA duplex. *J. Mol. Biol.*, **294**, 1215–1222.
- Chatake, T., Hikima, T., Ono, A., Ueno, Y., Matsuda, A., and Takénaka, A. (1999b). Crystallographic studies on damaged DNAs. II. N(6)-methoxyadenine can present two alternate faces for Watson–Crick base-pairing, leading to pyrimidine transition mutagenesis. *J. Mol. Biol.*, **294**, 1223–1230.
- Cheng, Y.-K. and Pettitt, B. M. (1992). Hoogsteen versus reversed-Hoogsteen base pairing: DNA triple helices. *J. Amer. Chem. Soc.*, **114**, 4465–4474.
- Chou, S.-H., Cheng, J.-W., and Reid, B. R. (1992). Solution structure of [d(ATGAGCGAATA)]₂: Adjacent G:A mismatches stabilized by cross-strand base-stacking and B_{II} phosphate groups. *J. Mol. Biol.*, **228**, 138–155.
- Chou, S.-H., Chin, K.-H., and Wang, A. H.-J. (2003). Unusual DNA duplex and hairpin motifs. *Nucleic Acids Res.*, **31**, 2461–2474.
- Chou, S.-H., Zhu, L., and Reid, B. R. (1997). Sheared purine x purine pairing in biology. *J. Mol. Biol.*, **267**, 1055–1067.
- Cmugelj, M., Sket, P., and Plavec, J. (2003). Small change in a G-rich sequence, a dramatic change in topology: new dimeric G-quadruplex folding motif with unique loop orientations. *J. Amer. Chem. Soc.*, **125**, 7866–7871.
- Coggio, S. and Xodo, L. E. (2006). G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription. *Nucleic Acids Res.*, **34**, 2536–2549.
- Dai, J., Punchihewa, C., Ambrus, A., Chen D., Jones, R. A., And Yang, D. (2007). Structure of the intramolecular human telomeric G-quadruplex in potassium solution: a novel adenine triple formation. *Nucleic Acids Res.*, **35**, 2440–2450.
- de Lange, T. (2005). Shelterin: the protein complex that shapes and safeguards human telomeres. *Genes Dev.*, **19**, 2100–2110.
- de los Santos, C., Rosen, M., and Patel, D. J. (1989). NMR studies of DNA (R⁺)_n•(Y⁻)_n•(Y⁺)_n triple helices in solution: imino and amino proton markers of T•A•T and C•G•C⁺ base-triple formation. *Biochemistry*, **28**, 7282–7289.
- Dexheimer, T. S., Sun, D., and Hurley, L. H. (2006). Deconvoluting the structural and drug-recognition complexity of the G-quadruplex-forming region upstream of the bcl-2 P1 promoter. *J. Amer. Chem. Soc.*, **128**, 5404–5415.
- Duckett, D. R., Murchie, A. I. H., Giraud-Panis, M.-J. E., Pöhler, J. R., and Lilley, D. M. J. (1995). Structure of the four-way DNA junction and its interaction with proteins. *Phil. Trans. Roy. Soc. Lond.*, **B347**, 27–36.
- Eddy, J. and Maizels, N. (2006). Gene function correlates with potential for G4 DNA formation in the human genome. *Nucleic Acids Res.*, **34**, 3887–3896.
- Egli, M. and Pallan, P. S. (2007). Insights from crystallographic studies into the structural and pairing properties of nucleic acid analogs and chemically modified DNA and RNA oligonucleotides. *Ann. Rev. Biophys. Biomol. Struct.*, **36**, 281–305.
- Egli, M., Pallan, P. S., Pattanayek, R., Wilds, C. J., Lubini, P., Minasov, G., Dobler, M., Leumann, C. J., and Eschenmoser, A. (2006). Crystal structure of homo-DNA and nature's choice of pentose over hexose in the genetic system. *J. Amer. Chem. Soc.*, **128**, 10847–10856.
- Eichman, B. F., Ortiz-Lombardia, M., Aymami, J., Coll, M., and Ho, P. S. (2002). The inherent properties of DNA four-way junctions: comparing the crystal structures of Holliday junctions. *J. Mol. Biol.*, **320**, 1037–1051.
- Eichman, B. F., Vargason, J. M., Mooers, B. H. M., and Ho, P. S. (2000). The Holliday junction in an inverted repeat DNA sequence: sequence effects on the structure of four-way junctions. *Proc. Natl. Acad. Sci. USA*, **97**, 3971–3976.
- Escudé, C., Nguyen, C. H., Kukreti, S., Janin, Y., Sun, J.-S., Bisagni, E., Garestier, T., and Hélène, C. (1998). Rational design of a triple helix-specific intercalating ligand. *Proc. Natl. Acad. Sci. USA*, **95**, 3591–3596.
- Escudé, C., Giovannangeli, C., Sun, J.-S., Lloyd, D. H., Chen, J. K., Gryaznov, S., Garestier, T., and Hélène, C. (1996). Stable triple helices formed by oligonucleotide N3'→P5' phosphoramidates inhibit transcription elongation. *Proc. Natl. Acad. Sci. USA*, **93**, 4365–4369.

- Faria, M., Wood, C. D., Perrouault, L., Nelson, J. S., Winter, A., White, M. R. H., Hélène, C., and Giovannangeli, C. (2000). Targeted inhibition of transcription elongation in cells mediated by triplex-forming oligonucleotides. *Proc. Natl. Acad. Sci. USA*, **97**, 3862–3867.
- Felsenfeld, G., Davies, D. R., and Rich, A. (1957). Formation of a three-stranded polynucleotide molecule. *J. Amer. Chem. Soc.*, **79**, 2023–2024.
- Gaddis, S. S., Wu, O., Thames, H. D., DiGiovanni, J., Walborg, E. F., MacLeod, M. C., and Vasquez, K. M. (2006). A web-based search engine for triplex-forming oligonucleotide target sequences. *Oligonucleotides*, **16**, 196–201.
- Gao, X. and Patel, D. J. (1988). G(syn)•A(anti) mismatch formation in DNA dodecamers at acidic pH: pH-dependent conformational transition of G•A mispairs detected by proton NMR. *J. Amer. Chem. Soc.*, **110**, 5178–5182.
- Gao, Y.-G., Robinson, H., Sanishvili, R., Joachimiak, A., and Wang, A. H.-J. (1999). Structure and recognition of sheared tandem G × A base pairs associated with human centromere DNA sequence at atomic resolution. *Biochemistry*, **38**, 16452–16460.
- Gehring, K., Leroy, J.-L., and Guéron, M. (1993). A tetrameric DNA structure with protonated cytosine•cytosine base pairs. *Nature*, **363**, 561–565.
- Gellert, M., Lipsett, M. N., and Davies, D. R. (1962). Helix formation by guanylic acid. *Proc. Natl. Acad. Sci. USA*, **48**, 2013–2018.
- Gilbert, D. E. and Feigon, J. (1999). Multistranded DNA structures. *Curr. Opin. Struct. Biol.*, **9**, 305–314.
- Ginell, S. L., Kuzmich, S., Jones, R. A., and Berman, H. M. (1990). Crystal and molecular structure of a DNA duplex containing the carcinogenic lesion O6-methylguanine. *Biochemistry*, **29**, 10461–10465.
- Ginell, S. L., Vojtechovsky, J., Gaffney, B., Jones, R. A., and Berman, H. M. (1994). Crystal structure of a mispaired dodecamer, d(CGAGAATTC(O6^{Me})GCG)₂, containing a carcinogenic O6-methylguanine. *Biochemistry*, **33**, 3487–3493.
- Giovannangeli, C. and Hélène, C. (2000). Triplex technology takes off. *Nature Biotechnol.*, **18**, 1245–1246.
- Gotfredsen, C. H., Schultze, P., and Feigon, J. (1998). Solution structure of an intramolecular pyrimidine-purine-pyrimidine triplex containing an RNA third strand. *J. Amer. Chem. Soc.*, **120**, 4281–4289.
- Gowers, D. M. and Fox, K. R. (1999). Towards mixed sequence recognition by triple helix formation. *Nucleic Acids Res.*, **27**, 1569–1577.
- Greene, K. L., Jones, R. L., Li, Y., Robinson, H., Wang, A. H.-J., and Wilson, W. D. (1994). Solution structure of a GA mismatch DNA sequence, d(CCATGAATGG)₂, determined by 2D NMR and structural refinement methods. *Biochemistry*, **33**, 1053–1062.
- Griffin, L. C. and Dervan, P. B. (1989). Recognition of thymine•adenine base pairs by guanine in a pyrimidine triple helix motif. *Science*, **245**, 967–971.
- Griffin, L. C., Kiessling, L. L., Beal, P. A., Gillespie, P., and Dervan, P. B. (1992). Recognition of all four base pairs of double-helical DNA by triple-helix formation: design of nonnatural deoxyribonucleosides for pyrimidine•purine base pair binding. *J. Amer. Chem. Soc.*, **114**, 7976–7982.
- Grigoriev, M., Praseuth, D., Guieysse, A. L., Robin, P., Thoung, N. T., Hélène, C., and Harel-Bellan, A. (1993). Inhibition of gene expression by triple helix-directed DNA cross-linking at specific sites. *Proc. Natl. Acad. Sci. USA*, **90**, 3501–3505.
- Haider, S., Parkinson, G. N., and Neidle, S. (2002). Crystal structure of the potassium form of an *Oxytricha nova* G-quadruplex. *J. Mol. Biol.*, **320**, 189–200.
- Hazel, P., Huppert, J., Balasubramanian, S., and Neidle, S. (2004). Loop-length-dependent folding of G-quadruplexes. *J. Amer. Chem. Soc.*, **126**, 16405–16415.
- Hazel, P., Parkinson, G. N., and Neidle, S. (2006). Topology variation and loop structural homology in crystal and simulated structures of a bimolecular DNA quadruplex. *J. Amer. Chem. Soc.*, **128**, 5480–5487.
- Hildebrand, B. A., Parel, S., and Leumann, C. J. (1997). 5-Substituted 2-aminopyridine *c*-nucleosides as protonated cytidine equivalents: increasing efficiency and selectivity in DNA triple-helix formation. *J. Amer. Chem. Soc.*, **119**, 5499–5511.
- Horvath, M. P. and Schultz, S. C. (2001). DNA G-quartets in a 1.86 Å resolution structure of an *Oxytricha nova* telomeric protein-DNA complex. *J. Mol. Biol.*, **310**, 367–377.
- Huppert, J. L. and Balasubramanian, S. (2005). Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.*, **33**, 2901–2907.
- Huppert, J. L. and Balasubramanian, S. (2007). G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res.*, **35**, 406–413.
- Hurley, L. H., Von Hoff, D. D., Siddiqui-Jain, A., and Yang, D. (2006). Drug targeting of the c-MYC promoter to repress gene expression via a G-quadruplex silencer element. *Seminars Oncol.*, **33**, 498–512.
- Jing, N., Zhu, Q., Yuan, P., Li, Y., Mao, L., and Tweardy, D. J. Targeting signal transducer and activator of transcription 3 with G-tetrad oligonucleotides: a potential novel therapy for head and neck cancer. (2006) *Mol. Cancer Ther.*, **5**, 279–286.

- Kalnik, M. W., Li, B. F. L., Swann, P. F., and Patel, D. J. (1989). O6-ethylguanine carcinogenic lesions in DNA: an NMR study of O6⁺G•C pairing in dodecanucleotide duplexes. *Biochemistry*, **28**, 6182–6192.
- Kang, C., Berger, I., Lockshin, C., Ratliff, R., Moysis, R., and Rich, A. (1994). Crystal structure of intercalated four-stranded d(C3T) at 1.4 Å resolution. *Proc. Natl. Acad. Sci. USA*, **91**, 11636–11640.
- Kang, C., Berger, I., Lockshin, C., Ratliff, R., Moysis, R., and Rich, A. (1995). Stable loop in the crystal structure of the intercalated four-stranded cytosine-rich metazoan telomere. *Proc. Natl. Acad. Sci. USA*, **92**, 3874–3878.
- Keppler, M. D., Neidle, S., and Fox, K. R. (2001). Stabilisation of TG- and AG-containing antiparallel DNA triplexes by triplex-binding ligands. *Nucleic Acids Res.*, **29**, 1935–1942.
- Kettani, A., Gorin, A., Majumdar, A., Hermann, T., Skripkin, E., Zhao, H., Jones, R., and Patel, D. J. (2000a). A dimeric DNA interface stabilized by stacked A•(G•G•G•G)•A hexads and coordinated monovalent cations. *J. Mol. Biol.*, **297**, 627–644.
- Kettani, A., Basu, G., Gorin, A., Majumdar, A., Hermann, T., Skripkin, E., and Patel, D. J. (2000b). A two-stranded template-based approach to G•(C-A) triad formation: designing novel structural elements into an existing DNA framework. *J. Mol. Biol.*, **301**, 129–146.
- Kiessling, L. L., Griffin, L. C., and Dervan, P. B. (1992). Flanking sequence effects within the pyrimidine triple-helix motif characterized by affinity cleaving. *Biochemistry*, **31**, 2829–2834.
- Kouchakjian, M., Bodepudi, V., Shibutani, S., Eisenberg, M., Johnson, F., Grollman, A. P., and Patel, D. J. (1991). NMR structural studies of the ionizing radiation adduct 7-hydro-8-oxodeoxyguanosine (8-oxo-7H-dG) opposite deoxyadenosine in a DNA duplex. 8-Oxo-7H-dG(syn).dA(anti) alignment at lesion site. *Biochemistry*, **30**, 1403–1412.
- Kuryavyi, V., Majumdar, A., Shallop, A., Chernichenko, N., Skripkin, E., Jones, R., and Patel, D. J. (2001). A double chain reversal loop and two diagonal loops define the architecture of unimolecular DNA quadruplex containing a pair of stacked G(syn)•G(syn)•G(anti)•G(anti) tetrads flanked by a G•(T-T) triad and a T•T•T triple. *J. Mol. Biol.*, **310**, 181–194.
- Lane, A. N., Jenkins, T. C., Brown, D. J. S., and Brown, T. (1991). N.m.r. determination of the solution conformation and dynamics of the A•G mismatch in the d(CGCAAATGGCG)₂ dodecamer. *Biochem. J.*, **279**, 269–281.
- Langridge, R. and Rich, A. (1963). Molecular structure of helical polycytidylic acid. *Nature*, **198**, 725–728.
- Laughton, C. A. and Neidle, S. (1992). Molecular dynamics simulation of the DNA triplex d(TC)₅•d(GA)₅•d(C+T)₅. *J. Mol. Biol.*, **223**, 519–529.
- Laughton, C. A. and Neidle, S. (1992). Prediction of the structure of the Y•R•R(′)-type DNA triple helix by molecular modelling. *Nucleic Acids Res.*, **20**, 6535–6541.
- Lavery, R., Zakrzewska, K., Sun, J.-S., and Harvey, S. C. (1992). A comprehensive classification of nucleic acid structural families based on strand direction and base pairing. *Nucleic Acids Res.*, **20**, 5011–5016.
- Lei, M., Podell, E. M., and Cech, T. R. (2004). Structure of human POT1 bound to telomeric single-stranded DNA provides a model for chromosome end-protection. *Nature Struct. Mol. Biol.*, **11**, 1223–1229.
- Leonard, G. A., Booth, E. D., and Brown, T. (1990). Structural and thermodynamic studies on the adenine.guanine mismatch in B-DNA. *Nucleic Acids Res.*, **18**, 5617–5623.
- Leonard, G. A., Thomson, J., Watson, W. P., and Brown, T. (1990). High-resolution structure of a mutagenic lesion in DNA. *Proc. Natl. Acad. Sci. USA*, **87**, 9573–9576.
- Leonard, G. A., Zhang, S., Peterson, M. R., Harrop, S. J., Helliwell, J. R., Cruse, W. B. T., Langlois d'Estaintot, Kennard, O., Brown, T., and Hunter, W. N. (1995). Self-association of a DNA loop creates a quadruplex: crystal structure of d(GCATGCT) at 1.8 Å resolution. *Structure*, **3**, 335–340.
- Li, J., Correia, J. J., Wang, L., Trent, J. O., and Chaires, J. B. (2005). Not so crystal clear: the structure of the human telomere G-quadruplex in solution differs from that present in a crystal. *Nucleic Acids Res.*, **33**, 4649–4659.
- Li, Y., Zon, G., and Wilson, W. D. (1991a). NMR and molecular modeling evidence for a G.A mismatch base pair in a purine-rich DNA duplex. *Proc. Natl. Acad. Sci. USA*, **88**, 26–30.
- Li, Y., Zon, G., and Wilson, W. D. (1991b). Thermodynamics of DNA duplexes with adjacent G.A mismatches. *Biochemistry*, **30**, 7566–7572.
- Lilley, D. M. J. and Norman, D. G. (1999). The Holliday junction is finally seen with crystal clarity. *Nature Struct. Biol.*, **6**, 897–899.
- Lipscomb, L. A., Peek, M. E., Morningstar, M. L., Verghis, S. M., Miller, E. M., Rich, A., Essigmann, J. M., and Williams, L. D. (1995). X-Ray structure of a DNA decamer containing 7,8-dihydro-8-oxoguanine. *Proc. Natl. Acad. Sci. USA*, **92**, 719–723.
- Liu, K., Sasisekharan, V., Miles, H. T., and Raghunathan, G. (1996). Structure of Py•Pu•Py DNA triple helices. Fourier transforms of fiber-type x-ray diffraction of single crystals. *Biopolymers*, **39**, 573–589.
- Luu, K.-N., Phan, A. T., Kuryavyi, V., Lacroix, L., and Patel, D. J. (2006). Structure of the human telomere in K⁺ solution: an intramolecular (3 + 1) G-quadruplex scaffold. *J. Amer. Chem. Soc.*, **128**, 9963–9970.

- Lynch, R. L., Liu, H., Gao, J., and Kool, E. T. (2006). Toward a designed, functioning genetic system with expanded-size base pairs: solution structure of the eight-base xDNA double helix. *J. Amer. Chem. Soc.*, **128**, 14704–14711.
- Macaya, R., Wang, E., Schultze, P., Sklenář, V., and Feigon, J. (1992). Proton nuclear magnetic resonance assignments and structural characterization of an intramolecular DNA triplex. *J. Mol. Biol.*, **225**, 755–773.
- Maizels, N. (2006). Dynamic roles for G4 DNA in the biology of eukaryotic cells. *Nature Struct. Mol. Biol.*, **13**, 1055–1059.
- McMicken, H. W., Bates, P. J., and Chen, Y. (2003). Antiproliferative activity of G-quartet-containing oligonucleotides generated by a novel single-stranded DNA expression system. *Cancer Gene Ther.*, **10**, 867–869.
- Menchise, V., De Simone, G., Tedeschi, T., Corradini, R., Sforza, S., Marchelli, R., Capasso, D., Saviano, M., and Pedone, C. (2003). Insights into peptide nucleic acid (PNA) structural features: the crystal structure of a D-lysine-based chiral PNA-DNA duplex. *Proc. Natl. Acad. Sci. USA*, **100**, 12021–12026.
- Mergny, J.-L., Phan, A. T., and Lacroix, L. (1998). Following G-quartet formation by UV-spectroscopy. *FEBS Letters*, **435**, 74–78.
- Mergny, J.-L., Sun, J.-S., Rougée, M., Montenay-Garestier, T., Barcelo, E., Chomilier, J., and Hélène, C. (1991). Sequence specificity in triple helix formation: experimental and theoretical studies of the effect of mismatches on triplex stability. *Biochemistry*, **30**, 9791–9798.
- Nielsen, P. E. (2004). PNA Technology. *Mol. Biotechnol.*, **26**, 233–248.
- Nielsen, P. E., Egholm, M., Berg, R. H., and Buchardt, O. (1991). Sequence-selective recognition of DNA by strand displacement with a thymine-substituted polyamide. *Science*, **254**, 1497–1500.
- Nowakowski, J., Shim, P. J., Prasad, G. S., Stout, C. D., and Joyce, G. F. (1999a). Crystal structure of an 82-nucleotide RNA-DNA complex formed by the 10–23 DNA enzyme. *Nature Struct. Biol.*, **6**, 151–156.
- Nowakowski, J., Shim, P. J., Stout, C. D., and Joyce, G. F. (1999b). Alternative conformations of a nucleic acid four-way junction. *J. Mol. Biol.*, **300**, 93–102.
- Ortiz-Lombardia, M., González, A., Eritja, R., Aymami, J., Azorin, F., and Coll, M. (1999). Crystal structure of a DNA Holliday junction. *Nature Struct. Biol.*, **6**, 913–917.
- Osborne, S. D., Powers, V. E. C., Rusling, D. A., Lack, O., Fox, K. R., and Brown, T. (2004). Selectivity and affinity of triplex-forming oligonucleotides containing 2'-aminoethoxy-5-(3-aminoprop-1-ynyl)uridine for recognizing AT base pairs in duplex DNA. *Nucleic Acids Res.*, **32**, 4439–4447.
- Parkinson, G. N., Lee, M. P. H., and Neidle, S. (2002). Crystal structure of parallel quadruplexes from human telomeric DNA. *Nature*, **417**, 876–880.
- Pasternack, L. B., Lin, S.-B., Chin, T.-M., Lin, W.-C., Huang, D.-H., and Kan, L.-S. (2002). Proton NMR studies of 5'-d-(TC)₃(CT)₃(AG)₃-3' – a paperclip triplex: the structural relevance of turns. *Biophys. J.*, **82**, 3170–3180.
- Paukstelis, P. J. (2006). Three-dimensional DNA crystals as molecular sieves. *J. Amer. Chem. Soc.*, **128**, 6794–6795.
- Paukstelis, P. J., Nowakowski, J., Birktoft, J. J., and Seeman, N. C. (2004). Crystal structure of a continuous three-dimensional DNA lattice. *Chem. Biol.*, **11**, 1119–1126.
- Patel, D. J. (1992). Covalent carcinogenic guanine-modified DNA lesions: solution structures of adducts and cross-links. *Curr. Opin. Struct. Biol.*, **2**, 345–353.
- Patel, D. J., Bouaziz, S., Kettant, A., and Wang, Y. (1999). In *Oxford Handbook of Nucleic Acid Structure* (Ed. Neidle, S), pp 389–453, Oxford.
- Patel, D. J., Shapiro, L., Kozlowski, S. A., Gaffney, B. L., and Jones, R. A. (1986). Structural studies of the O6meG.T interaction in the d(C-G-T-G-A-A-T-T-C-O6meG-C-G) duplex. *Biochemistry*, **25**, 1036–1042.
- Petersson, B., Nielsen, B. B., Rasmussen, H., Larsen, I. K., Gajhede, M., Nielsen, P. E., and Kastrop, J. S. (2005). Crystal structure of a partly self-complementary peptide nucleic acid (PNA) oligomer showing a duplex-triplex network. *J. Amer. Chem. Soc.*, **127**, 1424–1430.
- Phan, A. T., Guéron, M., and Leroy, J.-L. (2000). The solution structure and internal motions of a fragment of the cytidine-rich strand of the human telomere. *J. Mol. Biol.*, **299**, 123–144.
- Phan, A. T., Kuryavyi, V., Burge, S. E., Neidle, S., and Patel, D. J. (2007). Structure of an unprecedented G-quadruplex scaffold in the human c-kit promoter. *J. Amer. Chem. Soc.*, **129**, 4386–4392.
- Phan, A. T., Kuryavyi, V., Gaw, H. Y., and Patel, D. J. (2005). Small-molecule interaction with a five-guanine-tract G-quadruplex structure from the human MYC promoter. *Nature Chem. Biol.*, **1**, 167–173.
- Phan, A. T., Modi, Y. S., and Patel, D. J. (2004). Propeller-type parallel-stranded G-quadruplexes in the human c-myc promoter. *J. Amer. Chem. Soc.*, **126**, 8710–8716.
- Phan, A. T. and Patel, D. J. (2003). Two-repeat human telomeric d(TAGGGTTAGGGT) sequence forms interconverting parallel and antiparallel G-quadruplexes in solution: Distinct topologies, thermodynamic properties, and folding/unfolding kinetics. *J. Amer. Chem. Soc.*, **125**, 15021–15027.
- Phillips, K., Dauter, Z., Murchie, A. I. H., Lilley, D. M. J., and Luisi, B. J. (1997). The crystal structure of a parallel-stranded guanine tetraplex at 0.95 Å resolution. *J. Mol. Biol.*, **273**, 171–182.

- Privé, G. G., Heinemann, U., Chandrasegaran, S., Kan, L.-S., Kopka, M. L., and Dickerson, R. E. (1987). Helix geometry, hydration, and G•A mismatch in a B-DNA decamer. *Science*, **238**, 498–504.
- Radhakrishnan, I., Patel, D. J., Veal, J. M., and Gao, X. L. (1992). Solution conformation of a G•TA [guanosine•thymidine-5'-adenylic acid] triple in an intramolecular pyrimidine•purine•pyrimidine DNA triplex. *J. Amer. Chem. Soc.*, **114**, 6913–6915.
- Radhakrishnan, I. and Patel, D. J. (1994). DNA triplexes: solution structures, hydration sites, energetics, interactions, and function. *Biochemistry*, **33**, 11405–11416.
- Radhakrishnan, I., de los Santos, C., and Patel, D. J. (1991). Nuclear magnetic resonance structural studies of intramolecular purine•purine•pyrimidine DNA triplexes in solution. Base triple pairing alignments and strand direction. *J. Mol. Biol.*, **221**, 1403–1418.
- Rajagopal, P. and Feigon, J. (1989). Triple-strand formation in the homopurine:homopyrimidine DNA oligonucleotides d(G-A)₄ and d(T-C)₄. *Nature*, **339**, 637–640.
- Rankin, S., Reszka, A. P., Huppert, J., Zloh, M., Parkinson, G. N., Todd, A. K., Ladame, S., Balasubramanian, S., and Neidle, S. (2005). Putative DNA quadruplex formation within the human c-kit oncogene. *J. Amer. Chem. Soc.*, **127**, 10584–10589.
- Rasmussen, H., Kastrop, J. H., Nielsen, J. N., Nielsen, J. M., and Nielsen, P. E. (1997). Crystal structure of a peptide nucleic acid (PNA) duplex at 1.7 Å resolution. *Nature Struct. Biol.*, **4**, 98–101.
- Petersen, B., Nielsen, B. B., Rasmussen, H., Larsen, I. K., Gajhede, M., Nielsen, P. E., and Kastrop, J. S. (2005). Crystal structure of a partly self-complementary peptide nucleic acid (PNA) oligomer showing a duplex-triplex network. *J. Amer. Chem. Soc.*, **127**, 1424–1430.
- Rawal, P., Kummarsetti, V. B., Ravindran, J., Kumar, N., Halder, K., Sharma, R., Mukerji, M., Das, S. K., and Chowdhury, S. (2006). Genome-wide prediction of G4 DNA as regulatory motifs: role in Escherichia coli global regulation. *Genome Res.*, **16**, 644–655.
- Read, M. A. and Neidle, S. (2000). Structural characterization of a guanine-quadruplex ligand complex. *Biochemistry*, **39**, 13422–13432.
- Rhee, S., Han, Z., Liu, K., Miles, H. T., and Davies, D. R. (1999). Structure of a triple helical dna with a triplex-duplex junction. *Biochemistry*, **38**, 16810–16815.
- Rujan, I. N., Meleney, J. C., and Bolton, P. H. (2005). Vertebrate telomere repeat DNAs favor external loop propeller quadruplex structures in the presence of high concentrations of potassium. *Nucleic Acids Res.*, **33**, 2022–2031.
- Rusling, D. A., Powers, V. E. C., Ranasinghe, R. T., Wang, Y., Osborne, S. D., Brown, T., and Fox, K. R. (2005). Four base recognition by triplex-forming oligonucleotides at physiological pH. *Nucleic Acids Res.*, **33**, 3025–3032.
- Sanchez, A. M., Volk, D. E., Gorenstein, D. C., and Lloyd, R. S. (2003) Initiation of repair of A/G mismatches is modulated by sequence context. *DNA Repair*, **2**, 863–878.
- Schultze, P., Hud, N. V., Smith, F. W., and Feigon, J. (1999). The effect of sodium, potassium and ammonium ions on the conformation of the dimeric quadruplex formed by the *Oxytricha nova* telomere repeat oligonucleotide d(G₄)T(G₄)G(G₄). *Nucleic Acids Res.*, **27**, 3018–3028.
- Sen, D. and Gilbert, W. (1988). Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. *Nature*, **334**, 364–366.
- Šket, P., Črnugelj, M., and Plavec, J. (2003). Small change in a G-rich sequence, a dramatic change in topology: new dimeric G-quadruplex folding motif with unique loop orientations. *J. Amer. Chem. Soc.*, **125**, 7866–7871.
- Sundquist, W. I. and Klug, A. (1989). Telomeric DNA dimerizes by formation of guanine tetrads between hairpin loops. *Nature*, **342**, 825–829.
- Shields, G. C., Laughton, C. A., and Orozco, M. (1997). Molecular dynamics simulations of the d(T•A•T) triple helix. *J. Amer. Chem. Soc.*, **119**, 7463–7469.
- Siddiqui-Jain, A., Grand, C. L., Bearss, D. J., and Hurley, L. H. (2002) Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc. Natl. Acad. Sci. USA*, **99**, 11593–11598.
- Simonsson, T. (2001). G-quadruplex DNA structures – variations on a theme. *Biol. Chem.*, **382**, 621–628.
- Skelly, J. V., Edwards, K. J., Jenkins, T. C., and Neidle, S. (1993). Crystal structure of an oligonucleotide duplex containing G•G base pairs: influence of mispairing on DNA backbone conformation. *Proc. Natl. Acad. Sci. USA*, **89**, 804–808.
- Smith, F. W. and Feigon, J. (1992). Quadruplex structure of *Oxytricha* telomeric DNA oligonucleotides. *Nature*, **356**, 164–168.
- Somerville, L., Krynetski, E. Y., Krynetskaia, N. F., Beger, R. D., Zhang, W., Marhefka, C. A., Evans, W. E., and Kriwacki, R. W. (2003). Structure and dynamics of thioguanine-modified duplex DNA. *J. Biol. Chem.*, **278**, 1005–1011.
- Sorensen, J. J., Nielsen, J. T., and Petersen, M. (2004). Solution structure of a dsDNA:LNA triplex. *Nucleic Acids Res.*, **32**, 6078–6085.

- Špačková, N., Berger, I., and Šponer, J. (1999). Nanosecond molecular dynamics simulations of parallel and antiparallel guanine quadruplex DNA molecules. *J. Amer. Chem. Soc.*, **121**, 5519–5534.
- Špačková, N., Berger, I., and Šponer, J. (2001). Structural dynamics and cation interactions of DNA quadruplex molecules containing mixed guanine/cytosine quartets revealed by large-scale MD simulations. *J. Amer. Chem. Soc.*, **123**, 3295–3307.
- Strahan, G. D., Keniry, M. A., and Shafer, R. H. (1998). NMR structure refinement and dynamics of the K^+ -[d(G₃T₄G₃)]₂ quadruplex via particle mesh Ewald molecular dynamics simulations. *Biophys. J.*, **75**, 968–981.
- Strobel, S. A. and Dervan, P. B. (1990). Site-specific cleavage of a yeast chromosome by oligonucleotide-directed triple-helix formation. *Science*, **249**, 73–75.
- Sundquist, W. I. and Klug, A. (1989). Telomeric DNA dimerizes by formation of guanine tetrads between hairpin loops. *Nature*, **342**, 825–829.
- Tarköy, M., Phipps, A. K., Schultze, P., and Feigon, J. (1998) Solution structure of an intramolecular DNA triplex linked by hexakis (ethyleneglycol) units: d(AGAGAGAA-(EG)₆-TTCTCTCT-(EG)₆-TCTCTCTT). *Biochemistry*, **37**, 5810–5819.
- Thorpe, J. H., Gale, B. C., Teixeira, S. C., and Cardin, C. J. (2003). Conformational and hydration effects of site-selective sodium, calcium and strontium ion binding to the DNA Holliday junction structure d(TCGGTACCGA)₄. *J. Mol. Biol.*, **327**, 97–109.
- Todd, A. K., Johnston, M., and Neidle, S. (2005). Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res.*, **33**, 2901–2907.
- Uhlmann, E. (1998). Peptide nucleic acids (PNA) and PNA-DNA chimeras: from high binding affinity towards biological function. *Biol. Chem.*, **379**, 1045–1052.
- Van Meervelt, L., Vlieghe, D., Dautant, A., Gallois, B., Précigoux, G., and Kennard, O. (1995). High-resolution structure of a DNA helix forming (C•G)′G base triplets. *Nature*, **374**, 742–744.
- Vojtechovsky, J., Eaton, M. D., Gaffney, B., Jones, R. A., and Berman, H. M. (1995). Structure of a new crystal form of a DNA dodecamer containing T•(O6Me)G base pairs. *Biochemistry*, **34**, 16632–16640.
- Von Kitzing, E., Lilley, D. M. J., and Diekmann, S. (1990). The stereochemistry of a four-way DNA junction: a theoretical study. *Nucleic Acids Res.*, **18**, 2671–2683.
- Wang, Y. and Patel, D. J. (1992). Guanine residues in d(T₂AG₃) and d(T₂G₄) form parallel-stranded potassium cation stabilized G-quadruplexes with anti glycosidic torsion angles in solution. *Biochemistry*, **31**, 8112–8119.
- Wang, Y., de los Santos, C., Gao, X., Greene, K., Live, D., and Patel, D. J. (1991). Multinuclear nuclear magnetic resonance studies of Na cation-stabilized complex formed by d(G-G-T-T-T-T-C-G-G) in solution : Implications for G-tetrad structures. *J. Mol. Biol.*, **222**, 819–832.
- Wang, Y. and Patel, D. J. (1993). Solution structure of the human telomeric repeat d[AG₃(T₂AG₃)₃] G-tetraplex. *Structure*, **1**, 263–282.
- Wang, Y., Rusling, D. A., Powers, V. E. C., Lack, O., Osborne, S. D., Fox, K. R., and Brown, T. (2005). Stable recognition of TA interruptions by triplex forming oligonucleotides containing a novel nucleoside. *Biochemistry*, **19**, 5884–5892.
- Watson, J., Hays, F. A., and Ho, P. S. (2004). Definitions and analysis of DNA Holliday junction geometry. *Nucleic Acids Res.*, **32**, 3017–3027.
- Webster, G. D., Sanderson, M. R., Skelly, J. V., Neidle, S., Swann, P. F., Li, B. F., and Tickle, I. J. (1990). Crystal structure and sequence-dependent conformation of the A•G mispaired oligonucleotide d(CGCAAGCTGGCG). *Proc. Natl. Acad. Sci. USA*, **87**, 6693–6697.
- Weil, J., Min, T., Yang, C., Wang, S., Sutherland, C., Sinha, N., and Kang, C. (1999). Stabilization of the i-motif by intramolecular adenine-adenine-thymine base triple in the structure of d(ACCCT). *Acta Crystallogr.*, **D55**, 422–429.
- Westhof, E. (1992). Westhof's rule. *Nature*, **358**, 459–460.
- Wright, W.-E., Tesmer, V. M., Huffman, K. E., Levene, S. D., and Shay, J. W. (1997). Normal human chromosomes have long G-rich telomeric overhangs at one end. *Genes Dev.*, **11**, 2801–2809.
- Yang, W. (2006). Poor base stacking at DNA lesions may initiate recognition by many repair proteins. *DNA Repair*, **5**, 654–666.
- Ying, L., Green, J. J., Li, H., Klenerman, D., and Balasubramanian, S. (2003). Studies on the structure and dynamics of the human telomeric G quadruplex by single-molecule fluorescence resonance energy transfer. *Proc. Natl. Acad. Sci. USA*, **100**, 14629–14634.
- Zimmerman, S. B., Cohen, G. H., and Davies, D. R. (1975). X-ray fiber diffraction and model-building study of polyguanylic acid and polyinosinic acid. *J. Mol. Biol.*, **92**, 181–192.

Further Reading

Mispairing

- Brown, T. and Kennard, O. (1992). Structural basis of DNA mutagenesis. *Curr. Opin. Struct. Biol.*, **2**, 354–360.
- Hunter, W.N. (1992). Crystallographic studies of DNA containing mismatches, modified and unpaired bases. *Methods Enzymol.*, **211**, 221–231.
- Modrich, P. (1987). DNA mismatch correction. *Ann. Rev. Biochem.*, **56**, 435–466.
- Modrich, P. (2006). Mechanisms in eukaryotic mismatch repair. *J. Biol. Chem.*, **281**, 30305–30309.
- Singer, B. and Grunberger, D. (1983). *Molecular Biology of Mutagens and Carcinogens*, Plenum Press, New York.

Triplexes

- Fox, K. R. (2000). Targeting DNA with triplexes. *Curr. Med. Chem.*, **7**, 17–37.
- Fox, K. R. and Brown, T. (2005). An extra dimension in nucleic acid sequence recognition. *Quart. Rev. Biophys.*, **38**, 311–320.
- Moser, H. and Dervan, P. B. (1987). Sequence-specific cleavage of double helical DNA by triple helix formation. *Science*, **238**, 645–650.
- Neidle, S. (1997). Recent developments in triple-helix regulation of gene expression. *Anti-Cancer Drug Des.*, **12**, 433–442.
- Soyfer, V. N. and Potaman, V. N. (1996). *Triple Helical Nucleic Acids*, Springer-Verlag, New York.
- Thuong, N. T. and Hélène, C. (1993). Sequence-Specific Recognition and Modification of Double-Helical DNA by Oligonucleotides. *Ang. Chem., Int. Edit.*, **32**, 666–690.

Telomeres and Quadruplexes

- Burge, S., Parkinson, G. N., Hazel, P., Todd, A. K., and Neidle, S. (2006). Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res.*, **34**, 5402–5415.
- Cech, T. R. (2000). Life at the End of the chromosome: telomeres and telomerase. *Ang. Chem., Int. Edit.*, **39**, 34–43.
- Cech, T. R. (2004). Beginning to understand the end of the chromosome. *Cell*, **116**, 273–279.
- Davies, J. T. (2004). G-tetrads 40 years later: from 5''-GMP to molecular biology and supramolecular chemistry. *Angew. Chem. Intl. Edit.*, **43**, 668–698.
- Krauskopf, A. and Blackburn, E. H. (2000). Telomeres and their control. *Ann. Rev. Genetics*, **34**, 331–358.
- Neidle, S. and Balasubramanian, S. (eds.) (2006). *Quadruplex Nucleic Acids*, Royal Society of Chemistry, Cambridge, UK.
- Phan, A. T., Kuryavyi, V. and Patel, D. J. (2006) DNA architecture: from G to Z. *Curr. Opin. Struct. Biol.*, **16**, 1–11.
- Williamson, J. R. (1994). G-quartet structures in telomeric DNA. *Ann. Rev. Biophys.*, **23**, 703–730.

DNA Junctions

- Hays, F. A., Watson, J. and Ho, P. S. (2003) Caution! DNA crossing: crystal structures of Holliday junctions. *J. Biol. Chem.*, **278**, 49663–49666.
- Ho, P. S. and Eichman, B. F. (2001). The crystal structures of DNA Holliday junctions. *Curr. Opin. Struct. Biol.*, **11**, 302–308.
- Lilley, D. M. J. (2000). Structures of helical junctions in nucleic acids. *Quart. Rev. Biophys.*, **33**, 109–159.

Principles of Small Molecule-DNA Recognition

5.1 Introduction

The recognition of a DNA molecule by a small molecule or protein can be nonspecific, not choosing between any individual nucleotide, or nucleotide sequence. At the other extreme, it can be highly specific, solely recognizing a defined sequence within a gene or even a genome. Selective recognition can be performed very effectively by DNA itself, as has been seen in Chap. 4, with DNA–DNA recognition being governed in large part by base–base hydrogen-bonding. A single strand of oligonucleotide can recognize

- Sequences within the strand itself, folding back to form an intramolecular duplex, triplex, or quadruplex
- Another single strand to form a duplex or a bimolecular quadruplex
- A duplex to form a triplex
- Three other strands to form a tetramolecular quadruplex

There is an increasingly stringent sequence requirement on going from duplex to triplex to quadruplex, as detailed in Chap. 4. All of these utilize the hydrogen-bonding ability of DNA bases over and above Watson–Crick hydrogen-bonding. This is also the manner in which direct readout of DNA sequence information can be achieved by small molecules and proteins.

The task of DNA sequence recognition has as its prerequisite knowledge of the frequency of occurrence of a given DNA sequence within a defined length of DNA. This can be directly obtained from a DNA databank if the entire target sequence is known, which is now the case for the human genome, and for genomes in many other species, eukaryotic and prokaryotic (see, e.g., the genome web sites at www.sanger.ac.uk and www.genome.gov).

The frequency of occurrence of a given sequence can be readily estimated on the basis of assuming a random distribution of the four nucleotides, and of sequences generally. It is then possible to calculate the frequency of occurrence for a given size of a DNA recognition site. The reciprocal of the former is the number of unique sites for a given site size. The data in Table 5.1 suggests that the human genome, with a size of 3.2 billion nucleotides, requires a recognition site of 16–18 nucleotides in order to be reasonably certain of uniqueness, for example if a single sequence within

Table 5.1 The Statistics of DNA Sequence Selectivity, Based on a Random Distribution of Nucleotides

Site size	Frequency n , as $1:n$
2	10
3	32
4	136
5	512
6	2,080
7	8,196
10	524,800
12	8,390,656
15	536,870,912
16	2,147,516,416

a gene of therapeutic interest is to be targeted. It has been estimated that there are about 20,000 genes coded by the entire genome, corresponding to about 1–2% of the human genome length.

As of June 2007 this corresponds to 21,724 known genes (http://www.ensembl.org/Homo_sapiens/index.htm). An important caveat is that alternative splicing variants may increase this number two- or threefold, and there is now much evidence of significant variations between individuals. Thus the sequence recognition problem suggests that the size of site for unique recognition may be to the lower end of the 16–18 nucleotide range. However since the human and other gene sequences are now freely available, genome-wide searches should be made at the outset if sequence recognition targeting is being considered, in order to ascertain the actual uniqueness of a target site. Software such as that available at the Ensembl browser (<http://www.ensembl.org>), readily enable such searches to be performed. Importantly, many types of sequence pattern are not randomly distributed, and may turn out to be under- or overrepresented.

In order for direct readout to occur, an incoming molecule has to access the base pairs via a DNA groove. The orientation of this incoming group is important since hydrogen bonding to and from bases, and hence direct readout, are predominantly directional interactions. The edges of the bases in Watson–Crick duplex DNA have hydrogen-bonding potential (Fig. 5.1) that is only partially satisfied by base-pairing itself (Seeman, Rosenberg, and Rich, 1976), and therefore is available for recognition by ligands with hydrogen-bonding complementarity. This pattern of hydrogen-bond donation and acceptance has built-in sequence-dependency. Thus, in the major groove, a C•G Watson–Crick base pair has a pattern of (donor, acceptor, acceptor) hydrogen-bonding that is distinct from that for the reversed G•C base pair. The A•T and T•A base pairs have identical major groove donor/acceptor patterns, but the presence of the methyl group on thymine introduces an asymmetry in the groove that can, at least in principle, enable effective discrimination between these two base-pair sequences. On the other hand, patterns of hydrogen-bonding in the minor grooves of each pair of sequences are symmetric, so discrimination on this basis alone by molecules entering the minor groove is not straightforward.

It is often assumed that the minor groove in B-type DNA sequences is always narrow compared to the major groove. This has led to the view that it is unlikely that all three hydrogen bonds at the minor-groove edge of a C•G or G•C base pair or the two in A•T or T•A base pairs, are simultaneously available for direct hydrogen-bonding

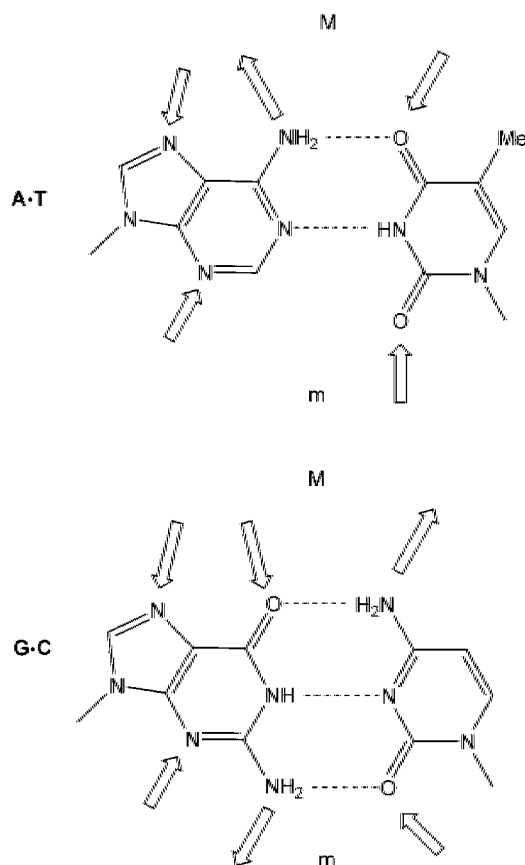


Figure 5.1 The pattern of hydrogen-bond donors and acceptors in A•T and G•C base pairs. The arrows indicate the directionality of the hydrogen bonds to and from the base edges. M and m indicate the major and minor grooves respectively.

sequence readout, for purely steric reasons. It has been established from the large number of oligonucleotide crystal structures that A/T-rich minor grooves tend to be narrower than average; their widths depend both on the length and the nature of the A/T sequence. Thus accessibility more generally is a consequence of groove-width variations, which are themselves sequence-dependent in both static and dynamic ways, as discussed in Chap. 3. However these trends are by no means absolute. A more realistic view is that A/T regions are more flexible than G/C ones, and that particular circumstances such as the requirements of the spine of hydration, often result in narrowed minor grooves in A/T tracts of sequence. Other circumstances where the energetics of interaction override the stability of structured water in a groove, can readily alter this feature so as to markedly widen the groove. Consequently, it is possible to simultaneously access the minor-groove edges of both components of the base pairs in a run of A/T sequence with appropriate ligands that can stabilize a widened groove. Other sequence-dependent features of base pairs and base-pair steps, which can be both localized (such as propeller and helical twist) and long-range (typified by the bending of A-tracts) will affect the geometric relationships between hydrogen-bond donors/acceptors on successive bases.

Indirect sequence readout by definition involves interaction with elements of DNA structure other than the base-pair hydrogen bonds, and is consequently inherently less directional. The backbones themselves provide several mechanisms for such readout. Their sequence-dependent structural features have been outlined in Chaps. 2 and 3. Advantage can be taken of differences in backbone and phosphate conformations such as B_I or B_{II} phosphate geometry, which can act as recognition signals. Perhaps more important are differences in distances between phosphate groups that necessarily result from backbone conformational differences, which can be recognized by hydrogen-bonding/electrostatic interactions with basic amino-acid side-chains. The B-DNA minor groove in particular has extended regions of apolar character. Carbon and hydrogen atoms of sugar residues, notably C1', H1', C4' H4', C5', and H5' form a significant part of the van der Waals and solvent-accessible surface walls of this groove (Fig. 5.2). The apolar groove surface is interspersed at regular intervals by highly polar phosphate groups. The surface of the groove floor comprises some carbon/hydrogen atoms along the base edges, together with the polar base hydrogen-bond donor/acceptor atoms. Groove depth is shallower in G/C regions on account of the exocyclic N2 substituent of guanine.

There are also significant differences in electronic character between A•T and G•C base pairs, with the latter being more electron-rich. Thus, electron-deficient planar molecules that can stack with individual base pairs (such as intercalating drug molecules, phenylalanine, or tryptophan side-chains), will tend to bind preferentially to G•C base pairs. Such preferences are by their nature localized to individual base-pair sites. There is a tendency for less stacking between bases in B-DNA duplexes for pyrimidine-3',5'-purine dinucleoside steps than purine-3',5'-pyrimidine ones (Fig. 5.3). The consequence of these differences is that the dinucleoside sequences CpG, TpG, CpA, and TpA (of pyrimidine-3',5'-purine type) are more readily unstacked than their counterparts GpC, GpT, ApC, and ApT. The cost in energetic terms between them is only a few kcal/mole. However this is sufficient for there to be a measurable preference for pyrimidine-3',5'-purine base steps in the first group to be more readily unstacked and separated so that planar molecules of appropriate size can be inserted and bind intercalatively between them.

The methyl group of thymine can play an important role in major-groove sequence readout, by being able to make stable van der Waals nonbonded contacts with hydrophobic groups or side-chains such as those of leucine, isoleucine, and alanine. Negative indirect readout can also take place when the presence of a particular group,

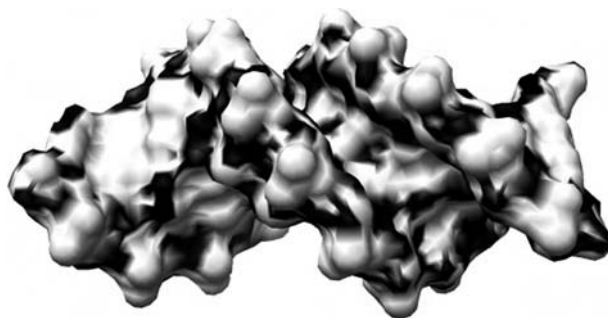


Figure 5.2 A surface representation of the Dickerson–Drew B-DNA duplex (PDB id 1BNA), looking onto the major and minor grooves, and with the hydrogen atoms at the surface of the grooves shaded black. Note that these form a significant part of the walls of the minor-groove surface.

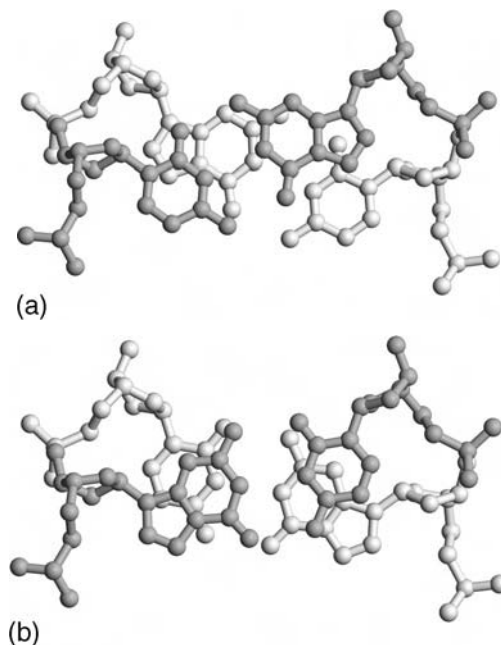


Figure 5.3 Base-stacking in a canonical B-DNA double helix at (a) pyrimidine-3',5'-purine dinucleoside steps and (b) purine-3',5'-pyrimidine steps.

substituent or side-chain would result in the occurrence of steric clashes with base substituents, and consequent destabilization of the resulting DNA-ligand complex. Examples include (i) the bulky methyl group of thymine driving major-groove recognition to a G/C region, and (ii) the bulky exocyclic amino group at the 2-position of guanine, whose presence in the minor groove acts as a deterrent to many minor-groove ligands. They then bind more readily in A/T regions.

The electronic characteristics of an individual base pair produce an electric field and potential in its vicinity. These electronic features are distinct for the various classical DNA structural types, as well as for changes in oligonucleotide sequences (Pullman and Pullman, 1981). Of special significance is the finding that for B-DNA an electrostatic potential minimum occurs in the major groove of oligo (dG)•(dC) sequences, and in the minor groove of oligo (dA)•(dT) sequences such as the 5'-AATT sequence in the Dickerson–Drew dodecamer. These all have greater negative electrostatic potentials than those around the phosphate groups, which explains why cationic ligands and basic protein side-chains tend to cluster in the grooves rather than around the phosphates. These electrostatic potentials and fields are important in directing an incoming ligand to a region of sequence; small local variations in net charge and field around an individual base pair then could assume some importance, especially for electrophilically driven covalent-bonding.

5.2 DNA-Water Interactions

Water molecules play a pivotal role in biological systems, in large part by use of the unique hydrogen-bonding ability of an individual water molecule. It can simultaneously

act as acceptor and donor for up to a total of four hydrogen bonds, often forming extended networks, even in bulk water. The hydration of DNA is essential for the maintenance of its structural integrity. At least 12–15 first-shell water molecules are associated with each nucleotide in the B-form. Historically, it has been presumed that some of these are clustered around phosphate groups, together with counter-ions such as sodium, potassium, or ammonium for charge neutralization. One would expect other water molecules to be associated with the bases, possibly taking advantage of the hydrogen-bonding facility (Fig. 5.1) offered by the edges of the bases (Seeman, Rosenberg, and Rich, 1976).

Fiber diffraction analyses have achieved some success in locating the positions of some counter-ions (Fuller, Forsyth, and Mahendrasingam, 2004). However only with the advent of medium- to high-resolution single-crystal analyses of oligonucleotides has it been possible to define the positions of the major proportion of the water molecules associated with a DNA sequence. NMR methods can provide valuable complementary information, by locating some of the least mobile water molecules, such as the minor-groove spine of hydration in the dodecamer d(CGCGAATTCGCG) (Kubinec and Wemmer, 1992; Liepinsh, Otting, and Wüthrich, 1992; Halle and Denisov, 1998). These methods can also provide data on residence times, which are typically in the low nanosecond range (Phan, Leroy, and Guéron, 1999). Crystallography, by contrast, provides a picture of hydration averaged over the timescale of the X-ray experiment, which even for a high-flux synchrotron radiation source, is many orders of magnitude greater than these times. Molecular dynamics simulations, which are routinely performed in the ns range, can thus realistically model water motions around oligonucleotides. A number of simulations have found excellent correlations between the majority of positions for statistical averaged waters from the simulations, and those found from crystallography (see, e.g., Duan *et al.*, 1997; Feig and Pettitt, 1999; Rueda *et al.*, 2004).

In general, the number of water molecules located by crystallography is highly dependent on the resolution of the structure and the thermal motion of individual water molecules (as well as on the quality of the raw X-ray diffraction data). It is necessary to achieve high resolution ($\geq 1.5 \text{ \AA}$) in a crystal structure analysis in order to find more than a fraction of the water molecules beyond the first hydration shell. The almost-universal use of cryo-crystallographic methods to collect low-temperature diffraction data, especially at high-intensity synchrotron X-ray sources, enables a significant proportion of these associated water molecules to be located, since the freezing of crystals at liquid nitrogen temperatures enables all molecular motions to be decreased, and also diminishes radiation damage. In particular the motions are reduced for quasi-bulk water molecules, which are involved in extensive water–water interactions (Pal *et al.*, 2003). In favorable cases complete second and even third-shell water molecules have been observed (see below). However some water molecules are undoubtedly disordered, and only partially occupy discrete positions. These are harder to locate and differentiate from background noise in electron-density maps. The availability of database search and comparison methods has enabled systematic studies of water clustering to be made using the large number of oligonucleotide structures in the Nucleic Acid Database (Schneider and Berman, 1995; Schneider, Patel, and Berman, 1998) (Table 5.2).

The nature and extent of DNA hydration also provide an indication of accessibility to ligands other than water itself, which are borne out by calculations of the solvent-accessible surface characteristics of DNA. These have shown (Alden and Kim, 1979) that there are some differences between A and B helical types in terms

Table 5.2 Some Oligonucleotide B-DNA Crystal Structures Showing Particular Hydration and Ion Features

Sequence	Feature	PDB ID code	NDB ID code
d(CGCGAATTCGCG)	Spine of hydration	2BNA	BDL002
d(CGCGAATTCGCG)	Spine of sodium ions	355D	BDL084
d(CGCGAATTCGCG)	Potassium ions	428D, 1FQ2	BD0005, BD0041
d(CGCGAATTCGCG)	Calcium ions	426D	BD0004
d(CCAGTACTGG)	Hydration at ultra-high resolution	1D8G	BD0023
d(CCATTAATGG)	Neutron diffraction study of hydration	1WQZ	-

of available surface area in the grooves, as well as between A•T and G•C base pairs. Sequence-dependent backbone conformational variability will also alter accessibility – when the minor groove in A/T regions is narrow its width is sufficient for the binding of only of one water molecule per base pair in the first hydration shell. The first and second shells together produce a network of water molecules, the spine of hydration (see below), which is necessarily two-dimensional in form.

The hydration patterns around phosphate groups in oligonucleotide crystal structures are dependent on whether the helix is an A, B, or Z type (Schneider, Cohen, and Berman, 1992; Saenger, Hunter, and Kennard, 1986). Detailed bridging hydration arrangements around phosphate groups have also been observed in fiber diffraction studies of DNA polymers (Fuller, Forsyth, and Mahendrasingam, 2004). Analysis of water molecules around phosphate groups in single crystals of A, B and Z-DNA duplexes has found that each of the two charged oxygen atoms in a phosphate group can accommodate up to three water molecules of hydration (Schneider, Patel, and Berman, 1998). The ester oxygen atoms, unsurprisingly, have little preference for hydration. In the A-form, successive phosphate groups along the phosphodiester backbone are close enough together for water molecules to be able to form bridges between them, as indeed is observed in many crystal structures. Such bridging patterns are not found in B-form oligonucleotides, with their more separated phosphate groups. These water molecules prefer to hydrogen bond to the anionic phosphate oxygen atoms, whereas the ester oxygen atoms O3' and O5' are rarely hydrated. By contrast the deoxysugar ring oxygen atom frequently participates in water networks, especially in A-form structures. Several A-DNA octamer high-resolution structures have extended major groove networks, which can form hexagonal, pentagonal, or ribbon-linked water molecules (Kennard *et al.*, 1986; Egli *et al.*, 1998). The coordination of these motifs to base edges appears to be a major factor in maintaining the integrity of A-form oligonucleotide structures.

DNA-bound water molecules are frequently associated with metal ions, especially sodium, calcium, potassium, or magnesium. These ions, together with ammonium, are required for neutralization of the negatively charged phosphate groups, although only a small proportion of such ions are typically located in crystal structures. It has been suggested that sodium ions can occupy part of the minor-groove spine of hydration in B-DNA, possibly in fast exchange with water molecules. This is discussed in detail below. Magnesium ions are normally hexa-hydrated, and have been well located in a number of oligonucleotide crystal structures. Typically, their role is to bridge phosphate groups (Fig. 5.4). They can also to lie in the major groove, contacting hydrogen-bond acceptors at the base edges (Fig. 5.5), in both instances with hydrogen bonds from the waters of hydration associated with the magnesium ion.

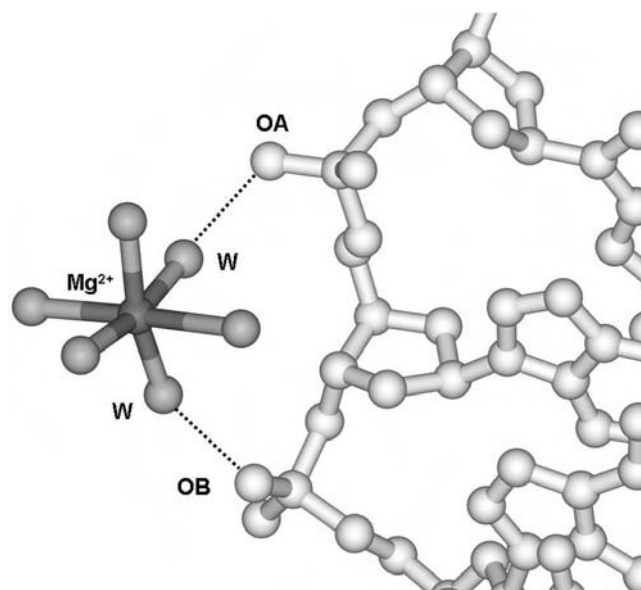


Figure 5.4 Part of the crystal structure of a typical B-DNA oligonucleotide, showing a hydrated magnesium ion coordinated to and bridging phosphate groups.

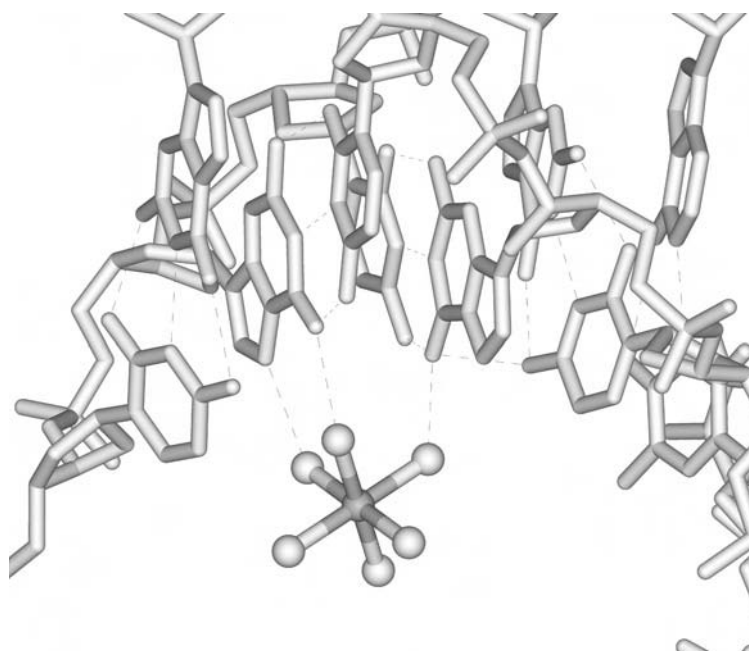


Figure 5.5 Another view from the same crystal structure, where symmetry relationships in the crystal also place the hydrated magnesium ion in the major groove, where it can coordinate to atoms on base edges.

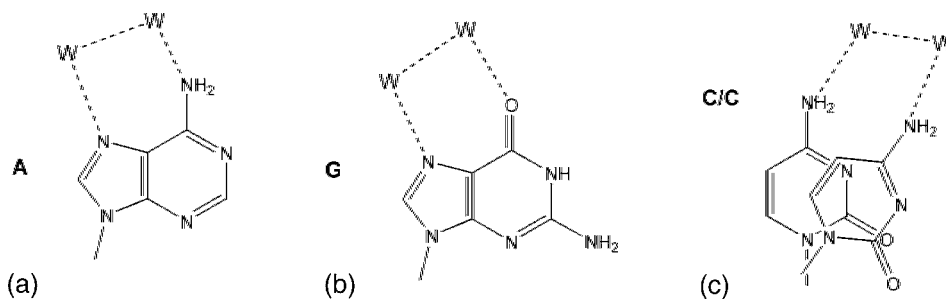


Figure 5.6 Typical patterns of water clusters in a B-DNA major groove: (a) and (b) are around individual purine bases, and (c) water molecules bridging between two consecutive pyrimidines in the same strand.

5.2.1 Hydration in the Grooves in Detail

Systematic examination of the observed distribution of water molecules in known oligonucleotide crystal structures has shown that the dependence of individual base hydration on the helical type (A, B, or Z), is ultimately due to the size and electronic characteristics of the grooves in each category. For example, the B-DNA major groove is dominated by large, hydrophobic thymine methyl groups in A/T regions. The width of this major groove is such that water-DNA networks would have to be three-dimensional, even for the first shell of hydration. These networks would need to straddle between and contact the two backbones as well as contacting base edge donor/acceptors. It appears that such arrangements are looser than minor groove ones, and associated water molecules are more mobile. It is thus unsurprising that water molecules have only been observed to form extensive major groove networks in B-DNA oligonucleotide structures studies at high resolution and undertaken at low temperatures. Instead, the major groove base hydration that has been observed in crystal structures is much more localized (Schneider and Berman, 1995). Water molecules tend to form discrete arrangements around particular bases, which sometimes extend to one or two bases in either direction. Purine bases frequently have water molecules hydrogen-bonded to them in the clustered manner shown in Fig. 5.6. Pyrimidine bases, by contrast, usually have a single associated water molecule, which can then hydrogen bond to another base, often via a second water molecule.

The hydration of the B-form minor groove has been the subject of much study and debate. The initial location (Drew and Dickerson, 1981) of the spine of hydration in the A/T region of the crystal structure of the sodium salt of the Dickerson–Drew duplex sequence $d(\text{CGCGAATTCGCG})_2$ was at a resolution of $1.9 \text{ \AA}$ (see Chap. 3), which is only moderate by current standards. Nonetheless, the pronounced network of solvent molecules observed in this analysis is still accepted as essentially correct. This network involves a linear array of hydrogen-bonded molecules extending over more than six base pairs in the central region of the minor groove, with every other solvent molecule also hydrogen-bonded to a pair of base-edge hydrogen-bond acceptors (Fig. 5.7). These thus bridge between adjacent base pairs.

The assignment of all these solvent atoms as water molecules has been reexamined in several low-temperature determinations of this structure, which also used crystallographic refinement methods based on more precise geometric parameters than were available to the original study. An analysis, extending to a resolution of $1.38 \text{ \AA}$,

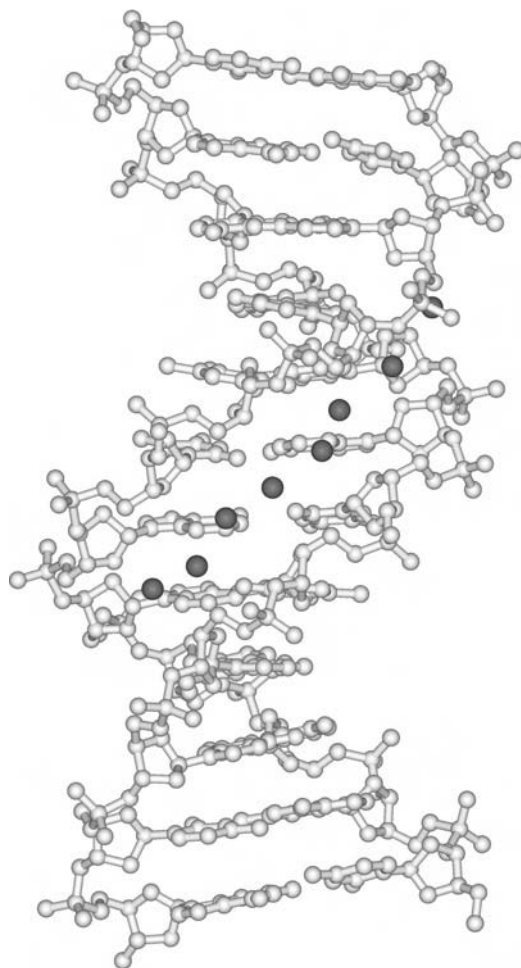


Figure 5.7 A plot of the spine of water molecules in the Dickerson–Drew dodecanucleotide crystal structure (Drew and Dickerson, 1981). The water molecules are shown as shaded spheres.

has suggested that the primary shell of solvent (which directly contacts base-pair edges), is partially occupied by sodium ions, and partly by water molecules (Shui *et al.*, 1998a). These assignments are based on the behavior of the solvent during the refinement, together with valence calculations. However subsequent crystal-structure determinations, one at yet higher resolution (1.1 Å) using synchrotron data, have not found unequivocal evidence for sodium-ion participation in the spine's structure (Tereshko, Minasov, and Efli, 1999; Chiu, Kaczor-Grzeskowiak, and Dickerson, 1999). The structure determinations of recently discovered crystal forms of the dodecamer also concur with this view (Liu and Subirana, 1999; Minasov, Tereshko, and Egli, 1999; Johansson, Parkinson, and Neidle, 2000). These analyses also show that minor-groove hydration extends well beyond the first and second shells of the spine itself, with each group of three water molecules forming the nucleus of a hexagon of higher-shell water molecules (Fig. 5.8). These eventually form hydrogen bonds to phosphate groups. The higher-shell solvent molecules in these hexagons are highly mobile, with large thermal factors. It is thus unsurprising that they have not been located in many studies.

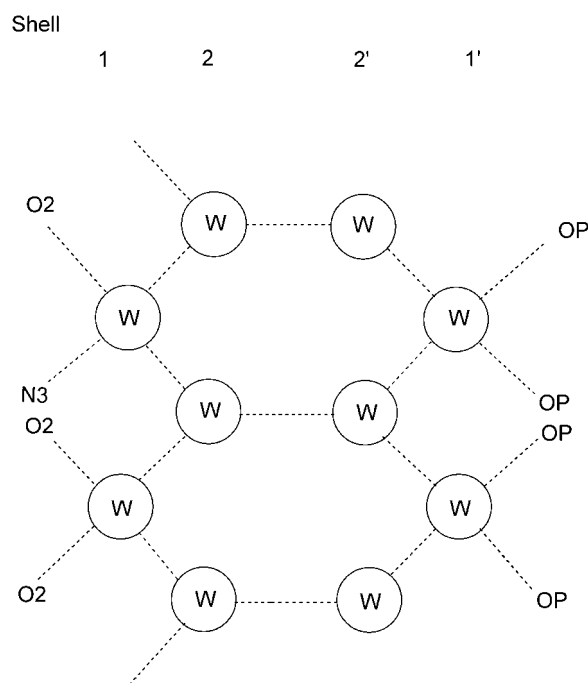


Figure 5.8 A schematic representation of the arrangement of higher-order shells of water molecules, as seen in high-resolution crystal structures of the Dickerson–Drew dodecanucleotide. The spine of hydration corresponds to shells 1 and 2.

Why has the unequivocal assignment of solvent molecules as waters or sodium ions been so difficult? Since X-rays are diffracted by electrons in atoms, there are obvious problems of differentiating one type of atom from another when the number of electrons in each is closely similar. An oxygen atom has eight electrons, and a sodium cation has ten. Even very high-resolution diffraction data has difficulty in distinguishing one from another. The problem is compounded by the mobility of solvent molecules, and by the probability that some have partial occupancies. On the other hand, discrimination between water molecules and potassium ions should be more straightforward with diffraction data extending to at least $2\text{ \AA}$. This is indeed the case (Shui *et al.*, 1998b). X-ray crystallography cannot hope to differentiate water molecules from ammonium ions. NMR approaches using isotopically labelled ^{15}N ammonium ions have been used to show that they also partially occupy some minor-groove sites, in fast exchange with water molecules (Hud, Sklenár, and Feigon, 1999). The overall picture that emerges is one in which small solvent molecules continually aggregate in and dissociate from various buried sites (the grooves) and around exterior sites (the phosphate groups), on the average in geometrically defined ways. Water molecules play the major role, but ammonium and alkali metal ions can also come into play. Both small molecule and protein–nucleic acid complexes are also dependent on solvent for their stability. Water molecules can play an active role in the recognition processes in these complexes even though they still have short residence times (Billiter *et al.*, 1996). Ions can play a significant role, in maintaining both local and large-scale conformational features such as A-tract bending (McConnell and Beveridge, 2000).

5.3 General Features of DNA-Drug and Small-Molecule Recognition

A large number of clinically important drugs and antibiotics are believed to exert their primary biological action by means of noncovalent interaction with DNA and subsequent inhibition of the template functions transcription and replication. Activity has been shown for many of these agents against a large number of human and animal diseases, ranging from cancer, bacterial infections, and a variety of viral diseases to parasitic diseases such as malaria and trypanosomiasis. Cancer chemotherapy has traditionally been in large part based on DNA-interacting cytotoxic drugs, a number of which are covalently binding agents, and many of these continue to be in clinical use today. Their effects are directly on DNA-template function. Thus, selective inhibition of DNA-directed RNA synthesis is produced by the antitumor antibiotic actinomycin D, whose noncovalent DNA interactions have been especially well-studied at the molecular level. For many of these molecules, drug-DNA interaction *in vivo* involves concomitant interactions with the enzymes DNA topoisomerase I or II to form ternary complexes. These can lead to lethal DNA double-strand breaks, and eventually to selective tumor cell death, often via activation of the p53 tumor suppressor protein apoptosis pathways. However cytotoxic drugs generally have high toxicity to both normal and cancer cells, and their therapeutic index (the ratio between a therapeutic and a toxic dose) is usually low. So unsurprisingly there has been a desire to rationally design new, improved ones, as well as to understand the molecular basis for the action of existing drugs. This has been the impetus for many of the large number of structural studies in this field. Almost without exception, these have been on binary drug-DNA complexes. Only recently has information become available at the atomic level on a few ternary complexes with protein partners. This is from crystallographic studies since the proteins involved, notably the topoisomerases, are generally too large for NMR structural analysis.

Present-day cancer (and anti-infective) therapeutics is mostly concerned with molecular targeted agents, directed against proteins, enzymes, and pathways that are selective to a particular cancer or infective agent. However DNA-interactive agents, even though they may appear to have the disadvantage of nonselectivity, in practice are often clinically useful because they target particular protein-DNA complexes that may be overexpressed in the disease state. For example, the utility of many DNA-reactive drugs in cancer is related to the overexpression of DNA topoisomerase enzymes associated with higher levels of proliferation in some cancers. The potential of DNA-interactive agents targeted against specific genes has yet to be fully realized; however their promise is such that their development into useful drugs continues to be an active area.

DNA-interactive small molecules can be conveniently categorized into two major classes based on their mode of interaction—noncovalent-binding and covalent-bonding. There are subdivisions within each class, which frequently overlap. Electrostatic-dominated noncovalent-binding involves sequence-neutral nonselective interactions between a cationic group and negatively charged phosphates on DNA. Examples of such molecules include the natural polyamines such as spermine and spermidine. They are frequently used in nucleic acid crystallization trials in order to shield the highly negatively charged individual DNA or RNA molecules from each other. It is rare for these polyamines to be located in electron density maps. This suggests that, in common with alkali metal ions, their positions are mobile. The majority of noncovalent-binding drugs carry formal positive charge, so electrostatic contributions are always a significant component of ligand-binding.

The rationalization of experimental drug-DNA structures by computer modeling methods has long been an active field, especially when such studies are accompanied by predictions of drug modifications with enhanced DNA affinity or selectivity. As with all theoretical studies on nucleic acids, there continues to be the problem of adequate models for electrostatics contributions to energetics, so that reliable quantitation of binding energies remains a challenge. Molecular dynamics simulations and free-energy calculations can correctly predict ranking order, for example, the superior binding of netropsin compared to distamycin (Dolenc *et al.*, 2005). Patterns of hydration in the minor groove are modified or displaced by drug-binding, and their analysis have been used to derive a set of potentials that have found use in quantitative docking studies (Ge, Schneider, and Olson, 2005). The ability to discriminate between different sites (i.e. sequences) and different ligands has assumed increasing importance with the widespread use of *in silico* methods to rapidly screen very large libraries of compounds in order to find leads for subsequent medicinal chemistry and pharmaceutical development. A popular and well-validated computer program for library screening is DOCK (see the web site at <http://dock.compbio.ucsf.edu/>). Calculation of ligand-DNA-binding energies within DOCK using the generalized-Born model for taking account of solvent-mediated electrostatic energetics, is now well-validated (Kang, Shafer, and Kuntz, 2004), and can enable a series of DNA-binding ligands to be reliably ranked. However scoring functions, even for the well-behaved ligand-minor groove complexes, are very sensitive to even very minor changes in geometry, and cannot always be relied upon to always correctly predict crystal structures of these complexes (Evans and Neidle, 2006). Docking algorithms aim to achieve a balance between accuracy and speed of calculation. If just one or a very few ligands are to be examined, then extensive molecular dynamics calculations can be employed (see, e.g., Cieplak *et al.*, 1990 for an early study on distamycin, and Dolenc *et al.*, 2005 for a more recent one). These enable conformational flexibility to be sampled to a greater extent than is possible during *in silico* library screening. Monte Carlo methods have also been used (Rohs *et al.*, 2005) in trial docking studies since they also offer these advantages over screening algorithms. However neither method enables whole libraries to be screened since these methods are still highly compute-intensive even for an individual drug-DNA system. None of these methods can as yet reliably predict absolute binding or free energies in a routine manner; the best that can be hoped for is accurate prediction of relative energies for a series of related molecules (Wang and Laughton, 2007).

5.4 Intercalative Binding

A large number of antibiotics, antibacterial, and antitumor drugs are characterized by the possession of an extended planar heteroaromatic ring system (chromophore). This essential structural feature is a chromophore of typically three fused six-membered rings (Fig. 5.9), which is approximately the same size as a base pair itself. The intercalation hypothesis, originally proposed by Leonard Lerman (Lerman, 1961) suggests that the planar chromophore of the drug molecule becoming inserted in between adjacent base pairs in an intercalative manner (Fig. 5.10). The drug chromophore is stabilized by polarization forces and van der Waals dispersion interactions between its planar group and the base pairs surrounding it. Optimal attractive interactions occur when the intercalating

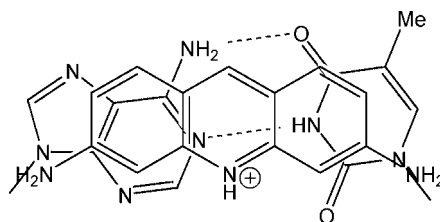


Figure 5.9 The proflavine molecule superimposed onto a base pair, showing their similar sizes.

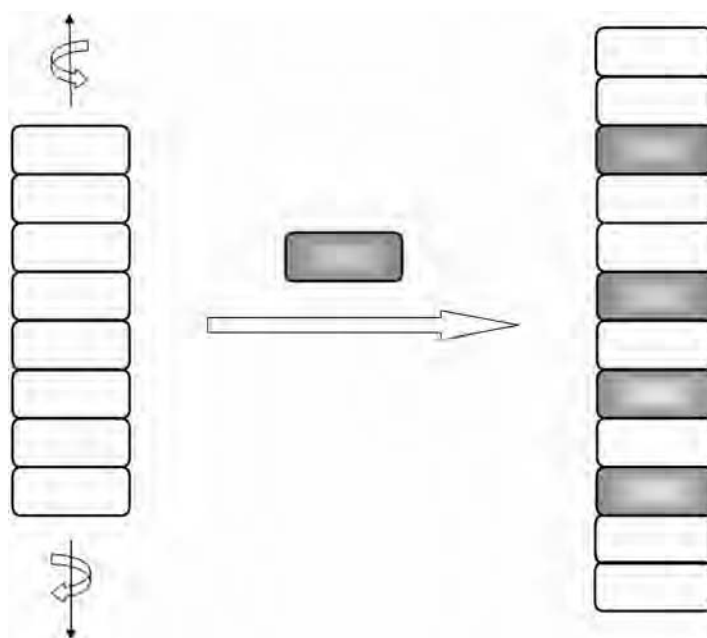


Figure 5.10 A schematic representation of a planar drug molecule binding intercalatively into a DNA double helix, causing base-pair extension and unwinding. Base pairs and drug molecules are shown respectively as open and shaded rectangles.

chromophore has polarized character, incorporating one or more hetero-atoms such as nitrogen. Electron-deficiency is also important, and a formal cationic charge is common (e.g. in the acridines). Intercalation results in an extension of the double helix by $3.4\text{ \AA}$ per bound drug molecule, together with changes in helical twist (unwinding) for the base pairs at and adjacent to each binding site (Waring, 1970). The base pairs either side of an intercalated drug thus increase in separation from $3.4\text{ \AA}$ to $6.8\text{ \AA}$. The maximum number of drug molecules which can be bound to a DNA duplex is one for every three base pairs, i.e. every other potential site becomes occupied at drug saturation. So the immediate base-pair step either side of a binding site cannot bind a drug molecule. This is the neighbor-exclusion principle, which originates from the interpretation of solution-binding data, but is fundamentally a consequence of the structural constraints forced on backbone geometry by intercalation.

Much of the interest in intercalating drugs has focused on their potential for anti-tumor activity. Some, such as the anthracycline antibiotic doxorubicin (also known as Adriamycin®) and its synthetic mimetic mitoxantrone, are still of major clinical value in the treatment of human cancers. For the majority of these drugs and their derivatives, DNA-binding abilities correlates approximately with biological activity. However, in spite of much activity in this field, the last 20 years has produced few new therapeutic agents with markedly superior activity to the long-established intercalative drugs. Drugs such as the anthracyclines form ternary complexes involving the topoisomerase family of proteins, which are often overexpressed in tumors. As a consequence they also produce DNA strand breaks, and can damage both tumor and normal cell DNA.

5.4.1 Simple Intercalators

DNA-intercalator recognition itself is essentially sequence-neutral, although, as described in Sect. 5.1, base-pair stacking/unstacking requirements usually result in a small preference for a pyrimidine-3',5'-purine sequence step at the binding site. "Simple intercalators" consist solely or primarily of an intercalating chromophore, often carrying a positive charge on the ring system, for example, proflavine and ethidium bromide (Fig. 5.11). The py-pu sequence preference at the drug-binding sites is generally obeyed both for these simple intercalators and for the complex ones such as daunomycin and its analogues. For the former, crystal structures of their complexes are restricted to DNA and RNA dinucleoside monophosphate "mini" duplexes, provided the sequences are pyrimidine-3',5'-purine (Neidle *et al.*, 1977). The complex intercalators, especially daunomycin and its analogues, have been mostly studied structurally (Fig. 5.12) as hexanucleotide complexes, with again drugs binding at the py-pu sites (see, e.g., Wang *et al.*, 1987). Analysis of conformational changes in intercalated dinucleoside duplex structures and the observation of "mixed" sugar puckers (C3'-*endo*-3',5'-C2'-*endo*), led initially to the suggestion that these are a necessary consequence of the intercalation process, and would explain the neighbor exclusion effect. This feature was not observed in several crystal structures involving the acridine drug proflavine (Neidle *et al.*, 1977), and is no longer considered to be indicative of any fundamental structural property. Final confirmation of this has come with analysis of the structure of a rhodium ligand-DNA complex (see below).

The degree of helix unwinding produced by intercalation is dependent on the nature of the bound drug. It can be measured experimentally (Waring, 1970) using covalently closed-circular DNA, with reference to the standard value of 26° per bound drug molecule, for ethidium bromide. Crystallographic analyses of intercalation complexes have shown qualitative unwinding; however the helical twist angle for the two base pairs immediately surrounding an intercalated drug, is not necessarily equivalent to this angle. The ribodinucleoside (CpG) duplex complex with proflavine (Neidle *et al.*, 1977) and the hexamer duplex d(CGTACG) with bound daunomycin (Wang *et al.*, 1987) both have normal B-DNA helical twist angles of 36° for the base pairs flanking the drug. In the latter case, there are changes in helical twist at adjacent base-pair steps. The unwinding angle is thus the sum of cumulative changes in helical twist over all affected base pairs. The process of base-pair separation to produce an intercalation site also results in a number of other changes, especially in backbone conformation, and sometimes also in the details of base-pair and base-step morphology. The crystal structure (Canals *et al.*, 2005) of the anticancer

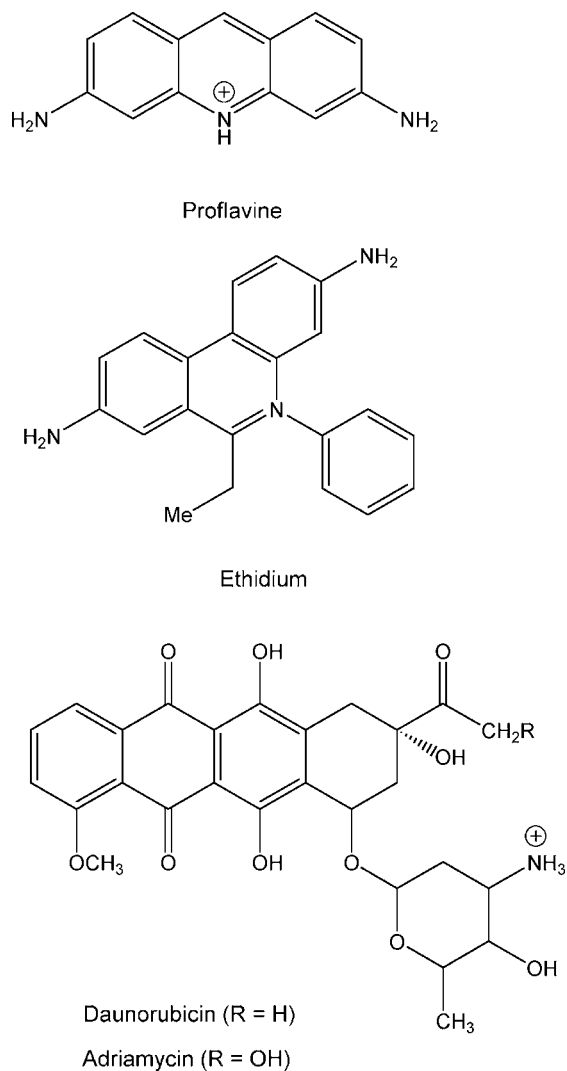


Figure 5.11 The structures of some representative intercalating molecules.

drug ellipticine (which has an experimentally determined unwinding angle of 22°) complexed with the hexanucleotide duplex formed by d(CGTACG), shows that the two drug-binding sites each have helical twist angles of $21\text{--}22^\circ$, with some further unwinding being apparent at adjacent sites. However this in total is still less than the experimental value, indicating the limitations of these crystallographic models for quantitatively mirroring intercalation into double-helical DNA.

5.4.2 Complex Intercalators

More complex intercalator molecules than ethidium or proflavine have attached groups such as side-chains, sugar rings, or peptide units, for example, the antitumor

drugs actinomycin D and daunomycin (Figs. 5.5, 5.12, 5.13). These groups normally reside in the minor groove of B-type DNA, where they are stabilized by van der Waals interactions, and can hydrogen-bond to adjacent bases, providing sequence-specific direct readout. For example, the threonine residue in the cyclic pentapeptide part of the actinomycin molecule hydrogen-bonds to the N2 and N3 atoms of



Figure 5.12 The crystal structure (Wang *et al.*, 1987) of an intercalation complex involving the anti-cancer drug daunomycin bound at the terminal CpG steps of the hexanucleotide duplex d(CGTACG). The two drug molecules are shown with shaded atoms, and the structure is viewed into the minor groove, where the daunosamine substituent sugar rings reside.

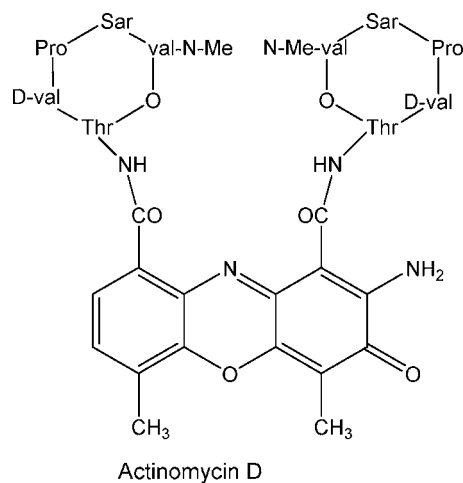


Figure 5.13 The structure of actinomycin.

a guanosine nucleoside, which is then structurally constrained to be on the 5'-side of the intercalating drug chromophore, giving a binding site requirement for the sequence GpX. This has been shown by a crystallographic analysis of actinomycin bound to the sequence d(GAAGCTTC). This structure shows that the phenoxazone chromophore of the drug is intercalated at the GpC site (Kamitori and Takusagawa, 1992) (Fig. 5.14). The two pentapeptide groups are situated in the minor groove of the DNA helix, in agreement with an NMR study of an actinomycin complex with the sequence d(AAAGCTTT) (Liu, Chen, and Patel, 1991). There have been a large number of crystallographic and NMR analyses of hexanucleotide complexes involving members of the anthracycline family of anticancer drugs and their analogues. Most adopt structures analogous to that first observed in the daunomycin complex (Wang *et al.*, 1987). More complex substituents than the monosaccharide of daunomycin and doxorubicin still reside in the minor groove, as shown in the crystal structure (Fig. 5.15) of the d(CGTACG) complex with the MAR70 analogue (Temperini *et al.*, 2005), which has two linked sugar group attached to the anthracycline chromophore. The crystal structure shows that these penetrate further into the minor groove on binding.

NMR and crystallographic studies on oligonucleotide complexes of the antitumor antibiotic nogalamycin (Fig. 5.16) have revealed a novel binding mode for this

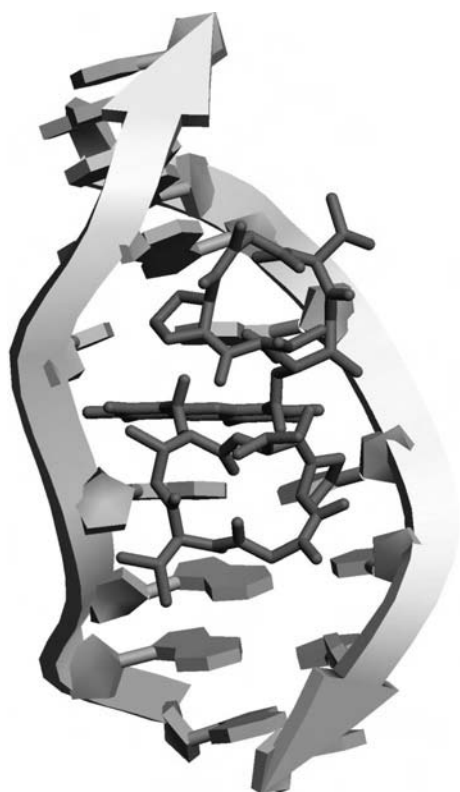


Figure 5.14 Crystal structure (Kamitori and Takusagawa, 1992) of a complex between actinomycin and the octanucleotide duplex formed by d(GAAGCTTC). The bonds of the drug molecule are shown as shaded sticks.

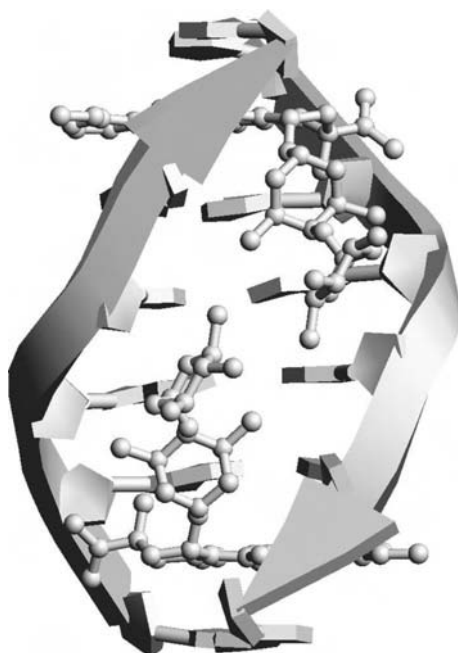


Figure 5.15 The crystal structure (Temperini *et al.*, 2005) of the anthracycline derivative MAR70 bound to the sequence d(CGTACG), showing the two saccharide rings of each drug molecule residing in the minor groove.

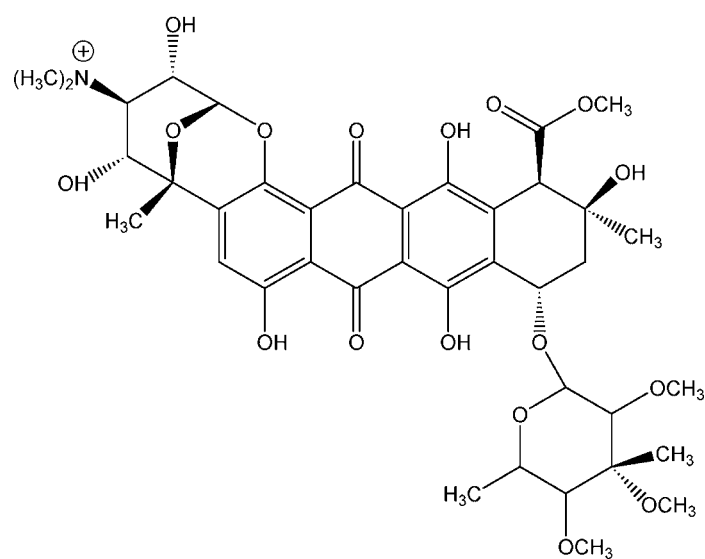


Figure 5.16 The structure of the antitumor antibiotic nogalamycin.



Figure 5.17 The crystal structure (PDB id 282D) of the complex between nogalamycin and the duplex sequence d(CCCGGG), with the drug molecule shown in ball-and-stick form.

drug, with the two attached groups residing, one in each groove; the drug chromophore itself is intercalated as expected (Egli *et al.*, 1991; Williams and Searle, 1999). The aglycone group is held in the minor groove and the amino sugar in the major groove (Fig. 5.17), with drug-oligonucleotide hydrogen-bonding to N2 and O6 atoms of guanines, thereby providing simultaneous major-and minor-groove direct sequence recognition, which agrees well with DNA footprinting data (Fox and Waring, 1986). In contrast with the simpler intercalators, the mechanism of nogalamycin intercalation into duplex DNA is not straightforward. The two attached groups, one at each end of the molecule, are sufficiently bulky to ensure that the drug chromophore cannot become intercalated without one or other of these groups effectively passing through or opening up the intercalation site. This implies that drug-DNA association and dissociation will both be exceptionally slow processes as a direct result of the structural constraints forced on the complex, as has been found experimentally.

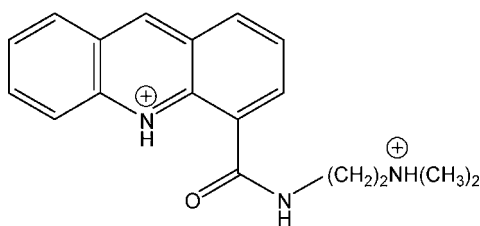
5.4.3 Major-Groove Intercalation

The majority of the “complex” intercalating molecules have their attached groups bound in the minor groove, with the DNA duplex retaining B-like character. These drugs tend to be natural products and the pendant groups are then sugar residues or modified peptides, often with hydrophobic character which complement the hydrophobic areas on the groove walls. By contrast, a number of intercalating molecules have been synthesized which subsequent structural studies have shown to have their attached groups bound in the major groove. It is plausible to speculate that these groups would be especially effective in competing with regulatory

Table 5.3 Crystal Structures of Some Drug-Oligonucleotide Intercalation Complexes

Drug	Sequence	PDB ID Code	NDB ID code
Proflavine	d(CG)	-	DDB009
Ditercalinium	d(CGCG)	1D32	DDD030
Ellipticine	d(CGTACG)	1Z3F	DD0070
Daunomycin	d(CGTACG)	1D11	DDF001
Daunomycin	d(CGATCG)	1DA0	DDF018
Adriamycin	d(CGATCG)	1D12	DDF019
4'-Epi-adriamycin	d(TGATCA)	1D54	DDF035
Bis-daunomycin	d(CGATCG)	1AGL	DDF072
Nogalamycin	d(CGTACG)	1D21	DDFA14
Nogalamycin	d(CCCGGG)	282D	DDF069
MAR70 disaccharide	d(CGTACG)	1R68	DD0061
9-Amino-DACA	d(CGATCG)	465D	DD0015
Triostin A	d(GCGTACGC)	1VS2	DDH010
Echinomycin	d(ACGTACGT)	2ADW	DD0073
Actinomycin	d(GAAGCTTC)	173D	DDH048
Rhodium complex	d(GUTGCAAC)	454D	UD0005
Rhodium complex	d(CGGAATTCCCG)	2O1I	DD0088
Cu-TMPy	d(CGATCG)	231D	DDF060
Ni-TMPy	d(CCTAGG)	1EM0	DD0027
Psoralen	d(CCGCTAGCGG)	1FHY	DD0030
Bis-acridine	d(CGATCG)	2GB9	DD0078

proteins which have motifs inserted into the major groove of their target DNAs. Major groove binding has been demonstrated for a family of synthetic anticancer drugs based on the acridine carboxamide skeleton. For example, the crystal structure has been determined (Todd *et al.*, 1999; Adams *et al.*, 2000) of a complex between the d(CGTACG)₂ duplex and the 9-amino derivative of the parent drug in this series of acridine derivatives, termed “DACA” (Fig. 5.18). This shows that the cationic end of the dimethylaminoethyl side-chain forms a hydrogen bond with O6 and N7 guanine base-edge atoms in the major groove (Fig. 5.19). Surprisingly, these drugs are also active against the enzyme DNA topoisomerase II, as are the minor-groove intercalators such as the anthracyclines daunorubicin, and Adriamycin. This suggests either that the active site of the enzyme does not discriminate between major- and minor-groove drug substituents on a DNA complex, or that there may be distinct subsites for the two categories. A related bis-acridine, also a 4-carboxamide derivative (Fig. 5.20) has been found, at least in the crystal lattice (Hopcroft *et al.*, 2006) to form a novel noncovalent interhelix cross-linked arrangement

**Figure 5.18** The structure of the acridine antitumour drug DACA.

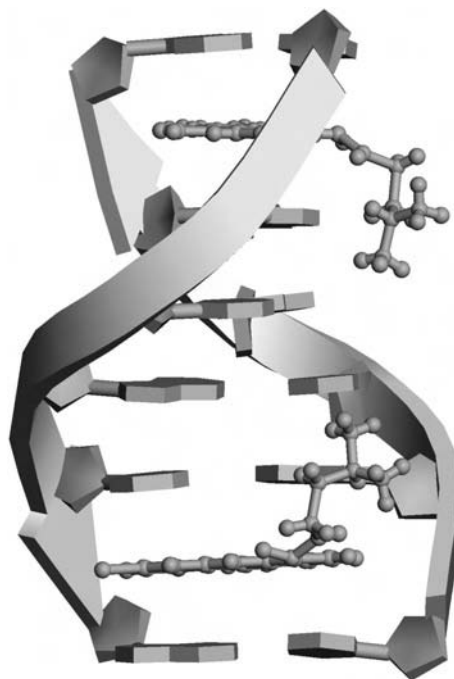


Figure 5.19 The crystal structure (Todd *et al.*, 1999) of a complex between 9-amino-DACA and the duplex sequence d(CGTACG), showing the side-chains in the major groove of the DNA and its hydrogen-bonding to base edges. The atoms of the 9-amino-DACA are shown shaded.

(Fig. 5.21), in which the two halves of the ligand each intercalate into a CG site, but in two separate symmetry-related duplexes.

A number of metal-containing intercalating molecules have been synthesized as probes of DNA structure, and several are photoactive. The rhodium complex (Fig. 5.22) is typical. It was designed to recognize the major groove of the sequence 5'-TGCA, i.e. intercalating unusually at the 5'-GC purine-3',5'-pyrimidine site. The remarkable crystal structure (Kielkopf *et al.*, 2000) of a complex with the recognition sequence embedded into an octanucleotide duplex has five independent complex molecules in the crystallographic asymmetric unit. They all show consistent structural features and conformations, with the planar group inserted from the major-groove direction, and the bulky nonplanar part of the ligand fitting snugly against the walls and floor of the major groove (Fig. 5.23). This high-resolution structure is one of the few for an intercalation complex where the ligand is embedded at the center of a significant length of DNA sequence, thus eliminating possible end-effects on conformation. It provides a view of the molecular features of intercalation that are relevant to the situation in bulk DNA. The long-standing controversy surrounding the nature of nucleoside sugar puckers at the intercalation site (see above) has been resolved by this structure determination, with the observation of consistent *C2'-endo* puckers at the ligand-binding site.

A series of rhodium complexes that have been designed to target mismatched DNA sequences, also interact by intercalative-type binding, as has been shown (Pierre,

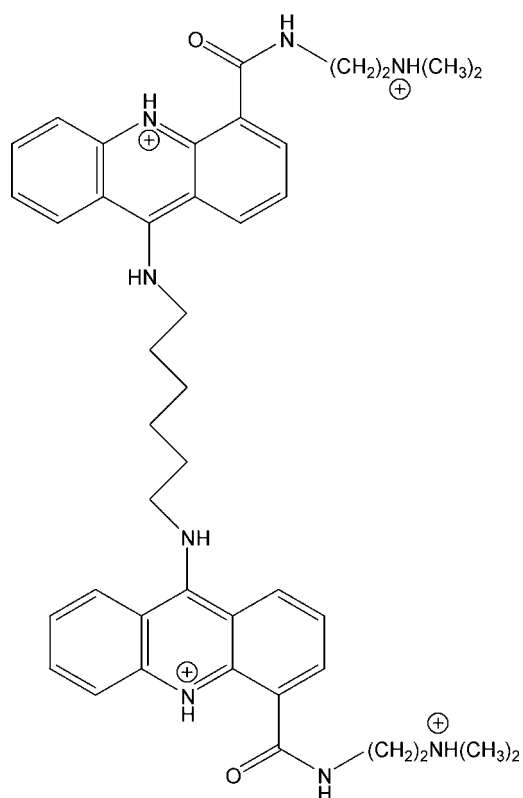


Figure 5.20 The structure of a bis-(4-carboxamide-acridine) molecule.

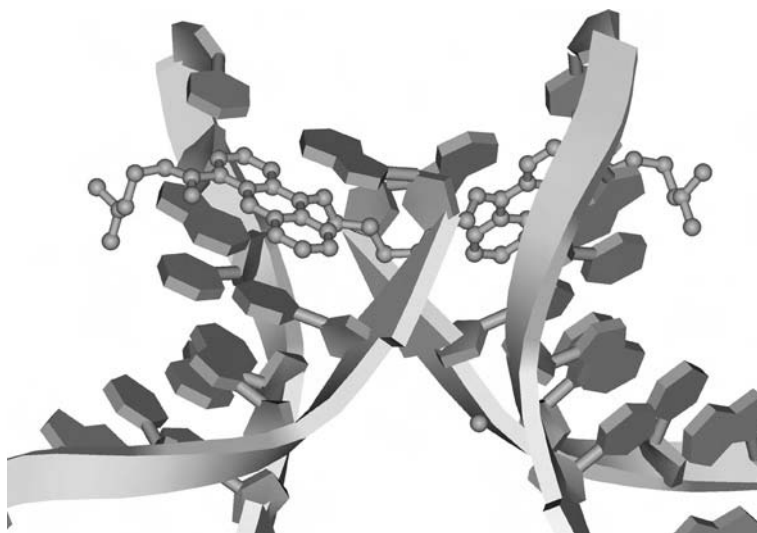


Figure 5.21 View of part of the crystal structure of the bis-acridine complex, showing the cross-linking arrangement that brings two duplexes close together (Hopcroft *et al.*, 2006).

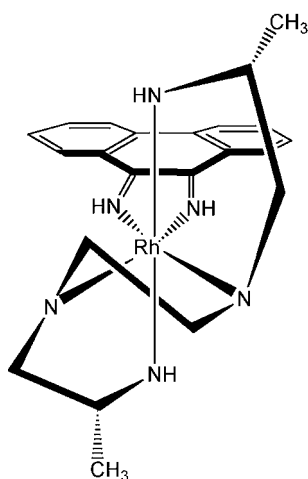


Figure 5.22 Structure of an intercalating rhodium complex (Kielkopf *et al.*, 2000).



Figure 5.23 Crystal structure of one duplex in the intercalation complex (Kielkopf *et al.*, 2000) between a rhodium compound and an octanucleotide duplex, with the rhodium complex shown bound between the central base pairs.

Kaiser, and Barton, 2007) in the crystal structure of a complex with a duplex formed by the sequence d(CGGAAATTCCCG). In this structure the bulky rhodium complex binds at two distinct sites. At a normal base-pair site, in the center of the helix, the ligand binds in a conventional intercalative manner (Fig. 5.24) and the two A•T base pairs separate by the standard 6.8 Å distance. However at the two A•C mismatch sites the ligand insertion results in the ejection of the adenine and cytosine bases, into the minor and major grooves respectively.

A highly unusual and unexpected intercalation motif has been observed with the hexamer sequence d(CGTACG) in complexes with several diverse drugs and ligands – an acridine-4-carboxamide (Adams *et al.*, 1999; Thorpe *et al.*, 2000), a bis-acridine, and a synthetic daunomycin analogue, ametantrone (Yang *et al.*, 2000). In each case, conventional major-groove intercalation into a duplex structure was expected. Instead, four

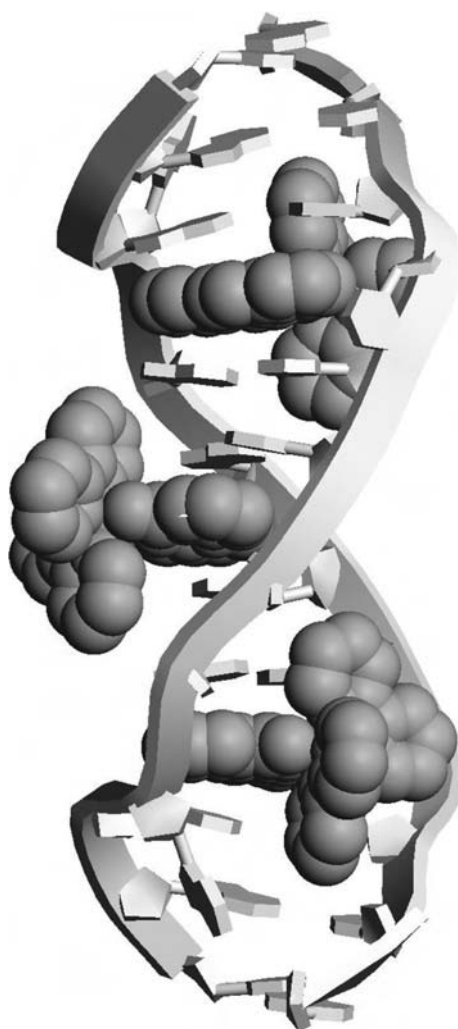


Figure 5.24 Crystal structure of a rhodium complex (shown in van der Waals space filling representation), bound to a mismatched DNA oligonucleotide (Pierre, Kaiser, and Barton, 2007).

duplexes in the crystal lattice come together to form an intercalation “platform” with terminal cytosines displaced out from one end of each duplex – this feature has been found in several native and protein-bound oligonucleotide crystal structures (see Chaps. 3 and 7). Each displaced cytosine then effectively strand-invades another symmetry-related duplex so that a C•G base pair is formed. A pseudo-intercalation cavity is formed by two such adjacent duplexes, which has some resemblance to four-way junctions, and to ligand-binding sites in DNA quadruplexes. Whether this type of structure has biological relevance, or is solely an artifact of crystal lattice packing, remains to be determined. The crystal structure of a true four-way junction complex has been determined with a covalently bound derivative of the largely planar molecule psoralen (Eichman *et al.*, 2001). Junction–ligand complexes are further discussed in Sect. 5.5.

Among more complex multigroove binding molecules are the porphyrins such as tetra-*N*-methyl-pyridyl-porphyrin (TMPyP4) (Fig. 5.25). Molecular modelling studies with this ligand have suggested that it could fit into a B-DNA intercalating site with two of its substituent *N*-methyl-pyridyl groups in each groove, although its observed sequence preferences have been interpreted in terms of minor-groove binding at A/T regions and intercalative in G/C ones. Intercalation has been inferred by NMR methods in an oligonucleotide complex to take place at the sequence CpG, in accordance with the pyrimidine-3',5'-purine sequence preference rule for intercalators (Marzilli *et al.*, 1986). This would necessarily place two of the nonplanar *N*-methylpyridine substituents in each groove, major and minor. However the crystal structure of a complex with the hexanucleotide duplex d(CGATCG)₂ and a copper complex of this porphyrin does not show full intercalation (Lipscomb *et al.*, 1996). Instead, the porphyrin is best described as being semi-intercalated at the two terminal CpG sites, with the end cytosine in each case being swung out of the helix (Fig. 5.26). Thus

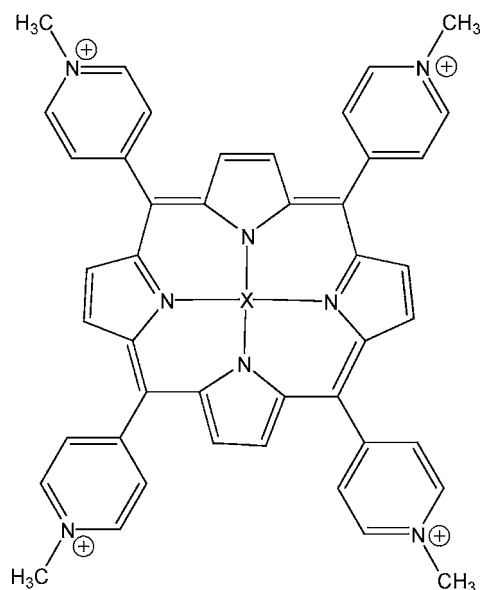


Figure 5.25 Structure of tetra-*N*-methylpyridine-porphyrin, TMPyP4. Atom X can be any one of a number of metal ions, such as nickel, copper, or zinc. The native porphyrin molecule has hydrogen atoms attached to two of the four inner-facing nitrogen atoms.

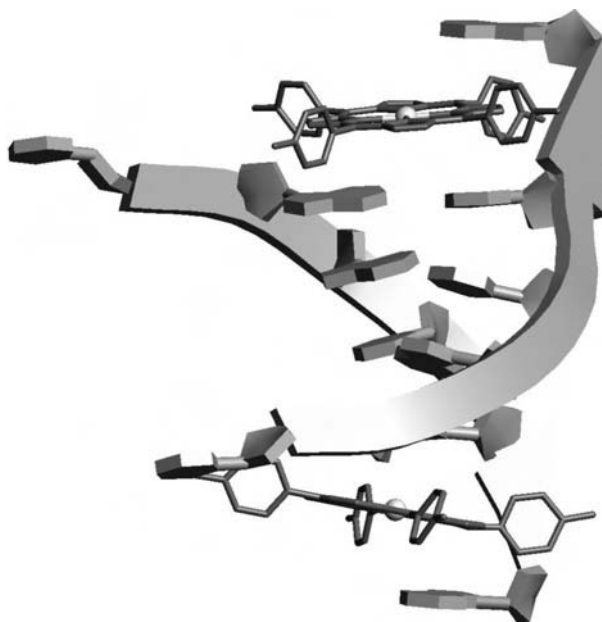


Figure 5.26 The crystal structure (Lipscomb *et al.*, 1996) of the copper complex of TMPyP4 bound to the sequence d(CGTACG). The two TMPyP4 molecules are shown in stick form, and the central coordinated copper atom is shown as a sphere.

the overall amount of porphyrin-base-stacking is less than expected for full intercalation. The nickel complex of the same porphyrin forms a quite distinct complex (Bennett *et al.*, 2000) with the d(CCTAGG)₂ duplex sequence. This hexanucleotide forms a conventional B-DNA duplex. However the porphyrin molecule, which is significantly buckled out of planarity, is not intercalated, but stacked at the ends of a pseudo-dodecanucleotide formed by two end-to-end hexanucleotide duplexes (Fig. 5.27). The crystal packing of this structure has one such dodecanucleotide duplex positioned with a porphyrin molecule capping the ends, and also positioned against the widened minor groove of another duplex, thereby providing a visualization of the groove-binding mode of this ligand.

5.4.4 Bis-Intercalators

Yet more complex are the drug molecules typified by the echinomycin and triostin family of antitumor antibiotics (Fig. 5.28), which have two intercalating chromophores linked together by cyclic oligopeptides. These molecules intercalate such that the two chromophores, whose planes are separated by a distance of 10.2 Å, can simultaneously bind at sites separated by two intervening base pairs (Fig. 5.29). This mode of binding, termed bis-intercalation, has been validated by crystallographic (Wang *et al.*, 1984; Quigley *et al.*, 1986) and NMR (Gao and Patel, 1988; Gilbert and Feigon, 1992) studies on echinomycin and triostin A bound to oligonucleotide sequences. Bis-intercalators can also be characterized by unwinding experiments on closed-circular

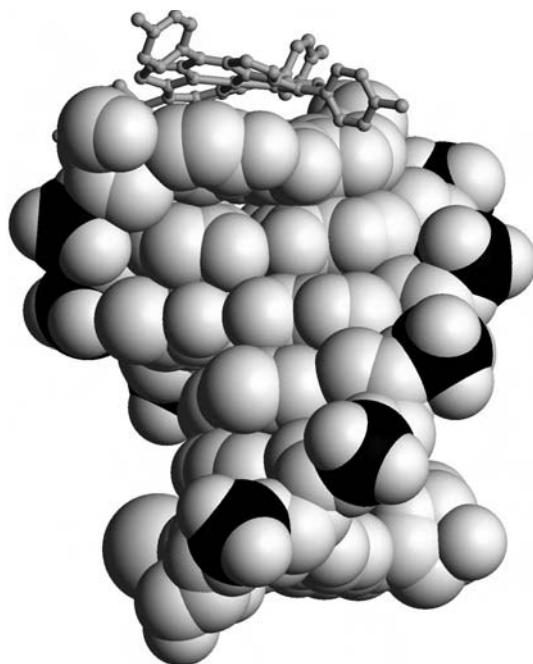


Figure 5.27 The crystal structure (Bennett *et al.*, 2000) of the nickel complex of TMPyP4 bound to the sequence d(CCTAGG). The TMPyP4 molecule is shown stacked at the top end of the B-form duplex.

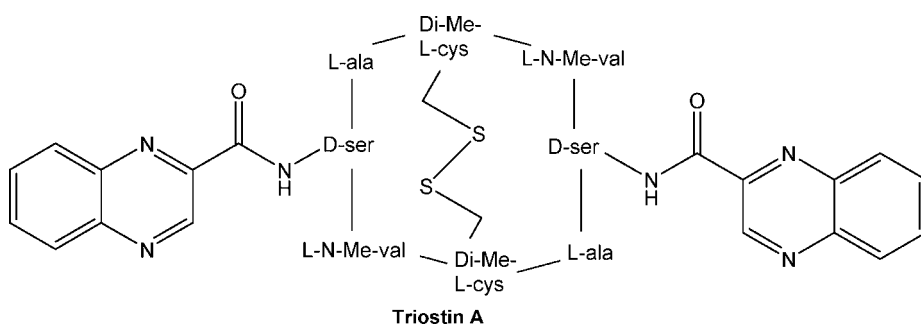


Figure 5.28 The structure of triostin A.

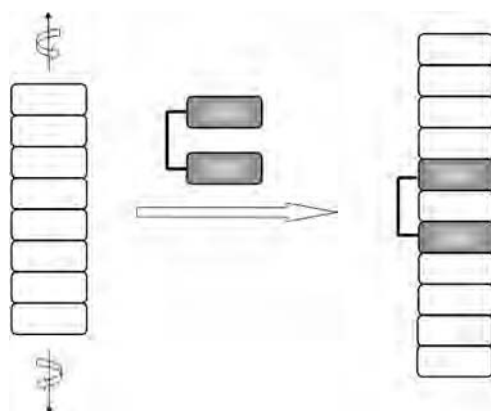


Figure 5.29 A schematic view of bis-intercalation.

DNA, since they produce approximately twice the unwinding per bound drug molecule compared to a mono-intercalator. Echinomycin itself has a sequence requirement for the dinucleoside 5'-CpG, with two C•G base pairs being flanked by the two quinoxaline chromophores. The peptide ring system sits in the DNA minor groove, with specific hydrogen-bonding to the 2-amino group of the guanine.

An unexpected finding in the crystal structures (Fig. 5.30) of the drug complexes has been the occurrence of Hoogsteen hydrogen-bonding for the base pairs “externally” flanking the bound drug – the two base pairs between the chromophores are always in a standard Watson–Crick arrangement. NMR studies indicate that this major structural change from standard B-DNA can occur with these drugs and certain short oligonucleotide sequences in solution. However chemical probe experiments (McLean, Seela, and Waring, 1989) with much longer (160 base-pair) sequences that are more truly representative of biological DNA, strongly suggest that Hoogsteen base-pairing does not occur in these sequences. It appears then that Hoogsteen-pairing does not always need to be formed as part of the structural changes in a DNA duplex consequent to bis-intercalative recognition, and that these changes

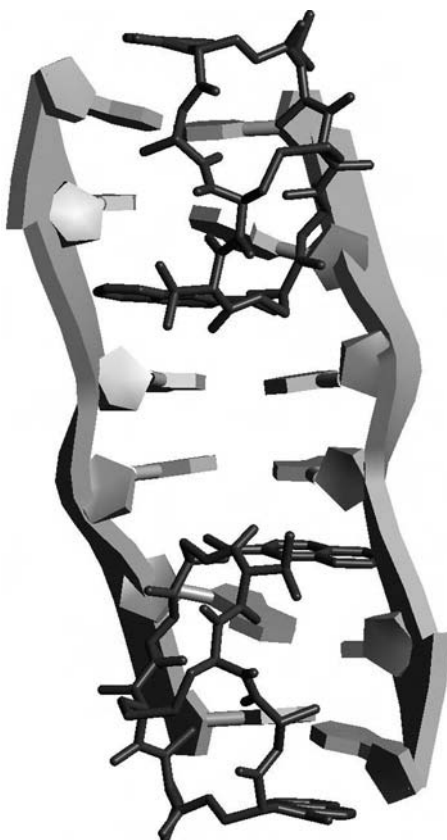


Figure 5.30 The crystal structure (Wang *et al.*, 1984) of a complex between an octanucleotide duplex and two triostin molecules, each bound in a bis-intercalative manner.

may be dependent on sequence length. A more recent higher-resolution (1.6 Å) crystal structure, has two crystallographically independent echinomycin complexes, each bis-intercalated into a duplex formed by the sequence d(ACGTACGT). This shows Watson–Crick base-pairing for three of the eight base pairs flanking the drug chromophore (Cuesta-Seijo, Weiss, and Sheldrick, 2006), whereas crystallographic studies on both this sequence in a different crystal form, and with a d(GCGTACGC) complex (Cuesta-Seijo and Sheldrick, 2005), showed a more consistent pattern of Hoogsteen base-pairing. This suggests that Hoogsteen pairing can occur, but that the energy barrier to the Watson–Crick arrangement is low and can even be overcome by crystal-packing forces.

A novel series of synthetic bis-intercalating analogues of daunomycin has been developed by the use of structure-based drug design principles (Chaires *et al.*, 1997), based on two assumptions:

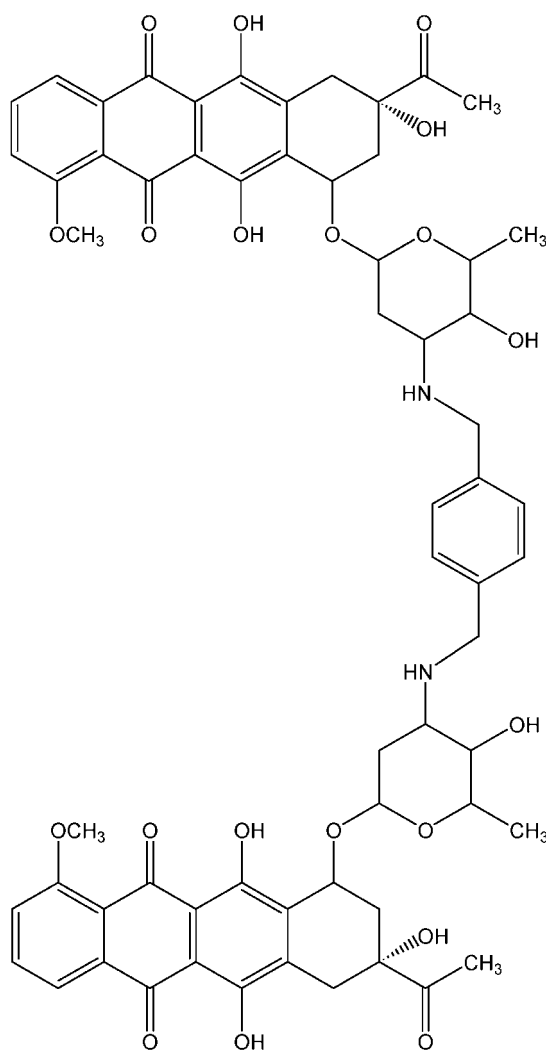


Figure 5.31 The structure of the bis-daunomycin molecule, WP631 (Chaires *et al.*, 1997).

- Such molecules would be superior DNA binders compared to daunomycin and its analogues since their two chromophores would together produce greater stacking interaction energy.
- Superior DNA affinity would result in superior biological activity, since it has been found in the anthracycline series of compounds that intercalative DNA-binding affinity correlates with cellular cytotoxicity. This in turn can indicate trends in antitumor activity. Thus it was anticipated that bis-daunomycins would show enhanced anticancer activity compared to their mono-analogues.

The well-established hexanucleotide-daunomycin crystal structures (Wang *et al.*, 1987), as described above, were taken as a starting-point, with two drug molecules bound in the hexanucleotide duplex. The two minor-groove pendant sugar groups were replaced by a p-xylene linker group through the exocyclic amino groups on each sugar ring, thus generating a single bis-intercalating molecule. The resulting molecule “WP631” (Fig. 5.31), has been shown to exhibit all of the enhanced DNA-binding properties predicted for a bis-intercalator. Both NMR and crystallographic studies (Fig. 5.32) have verified this binding mode (Hu *et al.*, 1997). The DNA affinity of WP631, at 10^{11} mole⁻¹, approaches that of many regulatory proteins. However this is not reflected in a quantum leap in biological response since its antitumor activity is not greatly superior to that of the native daunomycin. So, in spite of the elegance of the drug-design process in this instance, there appears to be an inherent limitation to the efficacy of intercalating molecules as antitumor agents. This is really unsurprising since even molecules such as WP631 will have high toxicity to normal cells, and are by themselves not capable of targeting specific oncogene sequences without at the same time binding to many other genes involved in normal cell functions.

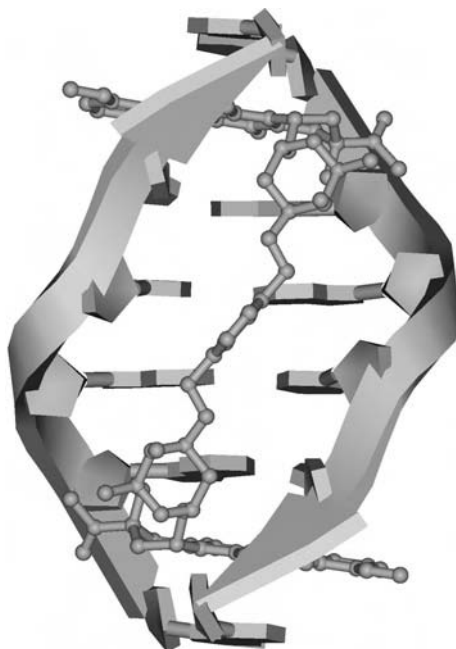


Figure 5.32 The crystal structure (Hu *et al.*, 1997) of the bis-daunomycin molecule WP631 bound to the duplex sequence d(CGTACG), viewed toward the minor groove and showing the linker group covering four base pairs.

Bis-intercalation from the major-groove direction is rare. It has been observed for bis-acridines (Hopcroft *et al.*, 2006; see above), and also (Peek *et al.*, 1994) in a molecule comprising two pyridocarbazole units, linked by a flexible aliphatic chain, which has been cocrystallized with the short duplex sequence d(CGCG). Each planar group of the ligand is intercalated at the CpG step, and the linker is positioned in the major groove. There is a marked bend in the DNA, toward the minor groove, which may be an innate feature induced by this type of bis-intercalating molecule.

5.5 Intercalative-Type Binding to Higher-Order DNAs

5.5.1 Triplex DNA–Ligand Interactions

Binding of small molecules to triplex DNA is a well-established approach for stabilizing tracts of triple-stranded sequences (Mergny *et al.*, 1992; Cassidy *et al.*, 1994; Fox *et al.*, 1995; Silver *et al.*, 1997; Escude *et al.*, 1998). The effective molecules in use have the common features of planar aromatic chromophores containing several fused electron-deficient rings, and it is presumed that they intercalate into the triplex at some point. A number show preferential binding to triplexes over duplex DNA. Effective stabilizing ligands include benzo[e]pyridoindole, the alkaloid coralyne, and a number of disubstituted amidoanthraquinones. There is little direct structural data on these complexes, although molecular modelling studies (e.g., with amidoanthraquinones: Keppler *et al.*, 1999) have successfully rationalized binding behavior in terms of intercalative-type models (Fig. 5.33). Selectivity for triple-stranded DNA arises from a combination of base-chromophore stacking and substituent binding in triplex grooves.

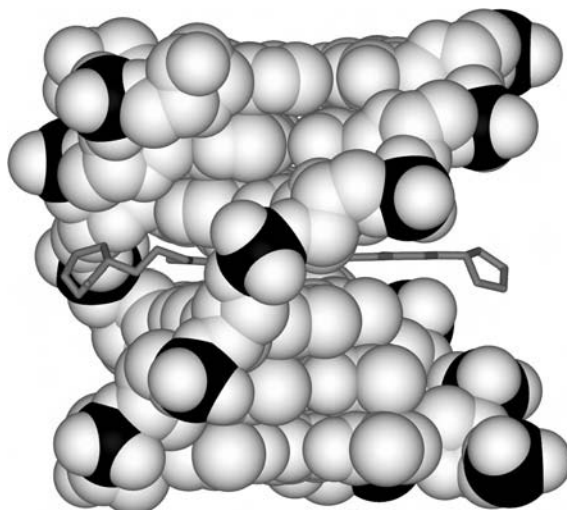


Figure 5.33 A molecular model of a disubstituted amidoanthraquinone molecule (shown in stick representation) intercalated into a B-type parallel triplex structure.

5.5.2 Ligand Binding to Quadruplex DNAs

The first report of drug-binding to a guanine-quadruplex type of DNA sequence involved the simple intercalator molecule ethidium (Guo *et al.*, 1992), although the mode of binding was not defined. This report lied dormant in the literature for some years, since when there has been an acceleration of interest in such four-stranded structures formed from telomeric and other DNA sequences, and their complexes with a range of intercalative type of molecules (Mergny and Hélène, 1998; Han and Hurley, 2000; Organisian and Bryan, 2007). This activity has arisen as a consequence of the findings that telomeric DNA is progressively shortened in normal cells, but is stabilized in length in most tumor cells by a specialized reverse transcriptase enzyme complex, telomerase (McEachern, Krauskopf, and Blackburn, 2000; Cech, 2000). This requires the telomere strand, acting as primer, to be single-stranded in order to hybridize with the RNA template component of the telomerase complex. Folding the primer strand into a quadruplex arrangement effectively inhibits telomerase (Zahler *et al.*, 1991). This folding process can be promoted by a range of intercalator-type molecules. A large number of compounds, including substituted anthraquinones, acridines, and the tetra-*N*-methyl-pyridyl porphyrin TMPyP4, all show correlations between quadruplex-binding and telomerase inhibition (see, e.g., Sun *et al.*, 1997; Wheelhouse *et al.*, 1998; Read *et al.*, 1999; Riou *et al.*, 2002). Many of the first generation of these molecules are also effective duplex-binding agents, and thus show only moderate selectivity for G-quadruplexes.

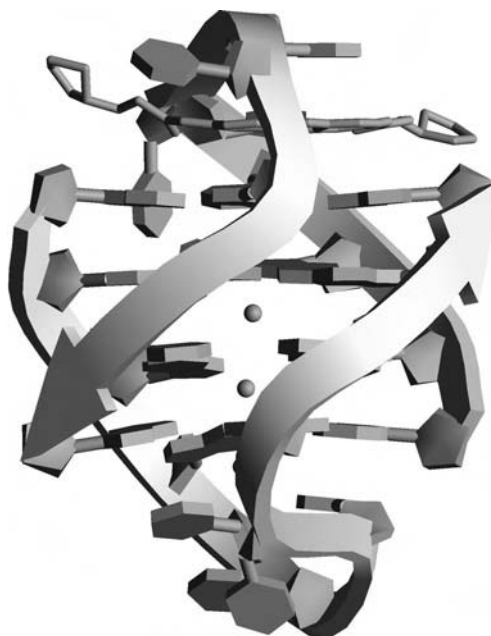
That it is possible to generate selective G-quadruplex-interactive molecules has been shown by several studies. Whereas ethidium itself binds only weakly to quadruplexes, its derivatives with appropriate side-chain functionality have been shown to bind with high affinity, and to be potent telomerase inhibitors, as have a number of dibenzophenanthroline derivatives (Koeppel *et al.*, 2001; Mergny *et al.*, 2001), the natural product telomestatin (Kim *et al.*, 2002) and trisubstituted acridine compounds (Table 5.4). These latter molecules have been successfully designed to exploit the groove differences between quadruplexes and duplexes, and are also more effective telomerase inhibitors (Read *et al.*, 2001).

This field is still in its infancy, with the diversity of possible quadruplex-type arrangements for G-rich (and C-rich) sequences providing a complex set of possible targets to be explored for drug action. As yet there are few experimental structures of ligand complexes, although most NMR, chemical cleavage, and molecular modelling studies conclude that intercalative-type binding does not occur between stacked G-quartets, but rather ligands are externally stacked at the ends of quadruplexes (Fedoroff *et al.*, 1998; Read and Neidle, 2000; Gavathiotis *et al.*, 2003). This binding-mode model is supported by the crystal structure (Haider, Parkinson, and Neidle, 2003) of a disubstituted amido acridine bound to a dimeric quadruplex formed by two molecules of the sequence d(GGGGTTTGGGG). The structure (Fig. 5.34) shows a single bound ligand molecule, at the end of four stacked G-tetrads constituting the quadruplex core. The ligand is held within the thymine loop, involving direct in-plane hydrogen-bonding between the acridine chromophore and a thymine base.

No detailed structures are available for the human four-repeat intramolecular quadruplex with bound drug, so the question of the preferred topology for this quadruplex is still unanswered. However the crystal structure (Parkinson, Ghosh, and Neidle, 2007) of a complex between TMPyP4 and a two-repeat bimolecular human telomeric quadruplex, shows that the all-parallel topology of the native bimolecular

Table 5.4 Crystal and NMR Structures of Drug Complexes with Higher-Order DNA Structures

Drug	Sequence	PDB ID Code	NDB ID code
Bis-acridine	d(TCGGTACCGA)	2GWA	DD0083
Bis-amido-acridine	d(GGGGTTTTGGGG)	1L1H	DD0049
TMPyP4	d(TAGGGTTAGGG)	2HRI	UD0071
TMPyP4	d(TGAGGGTGGIGAGGGTGGGGAAGG)	2A5R	-
Metallo-helicate	d(CGTACG)	2ET0	DD0077

**Figure 5.34** Crystal structure (Haider, Parkinson, and Neidle, 2003) of the 3,6-bis-[3-pyrrolidino-propionamide] acridine ligand complexed with the bimolecular quadruplex formed by the sequence d(GGGGTTTTGGGG). The ligand is shown in stick representation and the small spheres show the potassium ions in the central channel of the quadruplex structure.

quadruplex is retained, albeit with some changes to loop structure (Fig. 5.35). G-rich sequences are located both in telomeres, and in the promoter regions of many genes, especially those involved with cellular proliferation. These sequences, for example in the *c-myc* oncogene, also contain putative quadruplex-forming DNA, albeit in duplex form. *In vitro*, the G-rich strand of this *c-myc* promoter sequence forms an exceptionally stable quadruplex. The finding (Rangan, Fedoroff, and Hurley, 2001) of an association between the C-rich strand in the *c-myc* promoter and the quadruplex-binding ligand TMPyP4, shows that the i-motif may also be of significance as a drug target. A model for this complex has been suggested from NMR data, with the ligand bound externally to an i-motif tetraplex. An NMR structure (Phan *et al.*, 2005) of the *c-myc* quadruplex with bound TMPyP4 shows the ligand stacked externally onto this parallel-stranded quadruplex (Fig. 5.36). This confirms and extends the NMR structure for this porphyrin complexed with a G-quadruplex (Wheelhouse *et al.*, 1998), and both bear some resemblance to the crystal structure of nickel TMPyP4 stacked on the ends of a d(CCTAGG) duplex (Bennett *et al.*, 2000).

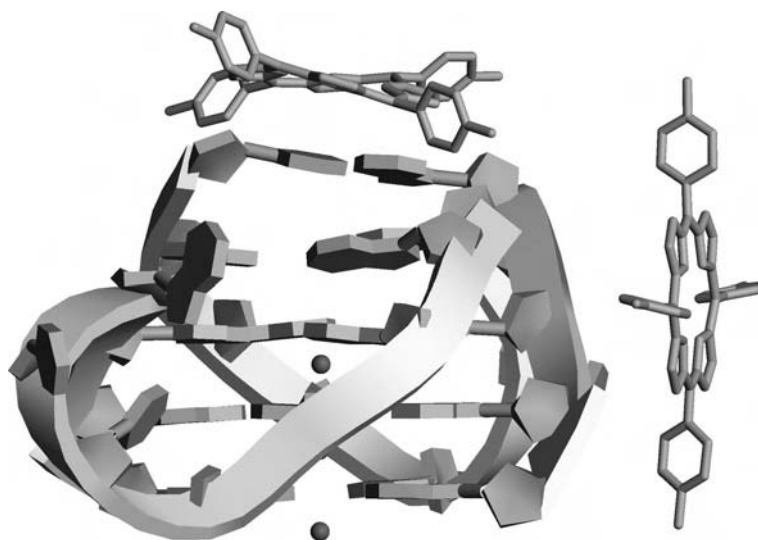


Figure 5.35 Crystal structure (Parkinson, Ghosh, and Neidle, 2007) of the bimolecular quadruplex formed by two strands of the human telomeric sequence d(TAGGGTTAGGG), as a complex with two TMPyP4 molecules. One is shown bound at the side of the quadruplex, stacking with a thymine residue, and the other is stacked on top of a set of base pairs above the G-tetrads.

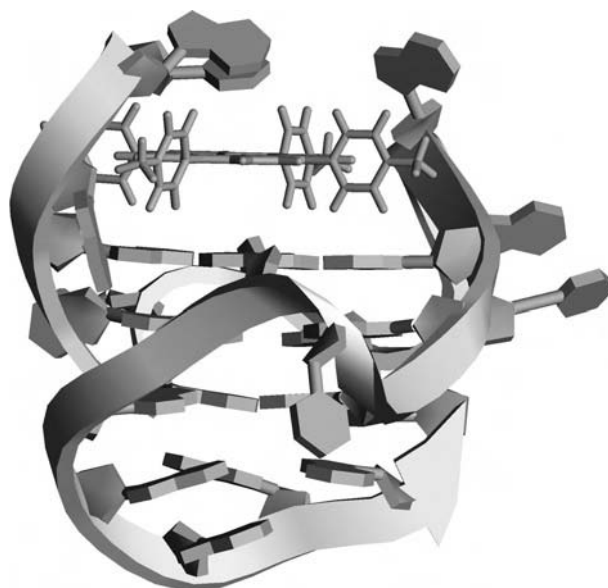


Figure 5.36 An NMR structure of the complex between the *c-myc* intermolecular quadruplex and TMPyP4 (Phan *et al.*, 2005). The TMPyP4 molecule is situated on the exterior of the G-tetrad stack.

5.5.3 Ligand Binding to Junction DNAs

The crystal structures of two DNA four-way Holliday junctions complexed with a derivative of the covalently binding intercalating drug psoralen (Eichman *et al.*, 2001), are remarkable demonstrations of the power of particular DNA sequences to form

non-duplex structures. One structure, with the sequence d(CCGGTACCGG), is not unexpected, given that this sequence forms a junction structure in the absence of ligand – the complex is nearly identical to the native structure. By contrast, the related sequence d(CCGCTAGCGG) dimerizes to a duplex in the absence of psoralen, but is induced by the drug to form a related (though nonidentical) junction structure. It has been suggested that this is a model to explain the promotion of recombination events by psoralen, leading to repair of the resulting lesions in DNA. These structures have raised the intriguing question of whether other, noncovalently binding drugs can similarly promote and stabilize four-way junctions, since the central junction regions in them can have considerable free space.

An answer to this question has come from the crystal structure of a Holliday junction with a bis-acridine molecule (Fig. 5.20), which is unexpectedly isomorphous with native Holliday junction structures (Brogden *et al.*, 2007). The drug was observed in clear electron density bound at the central crossover region (Fig. 5.37);

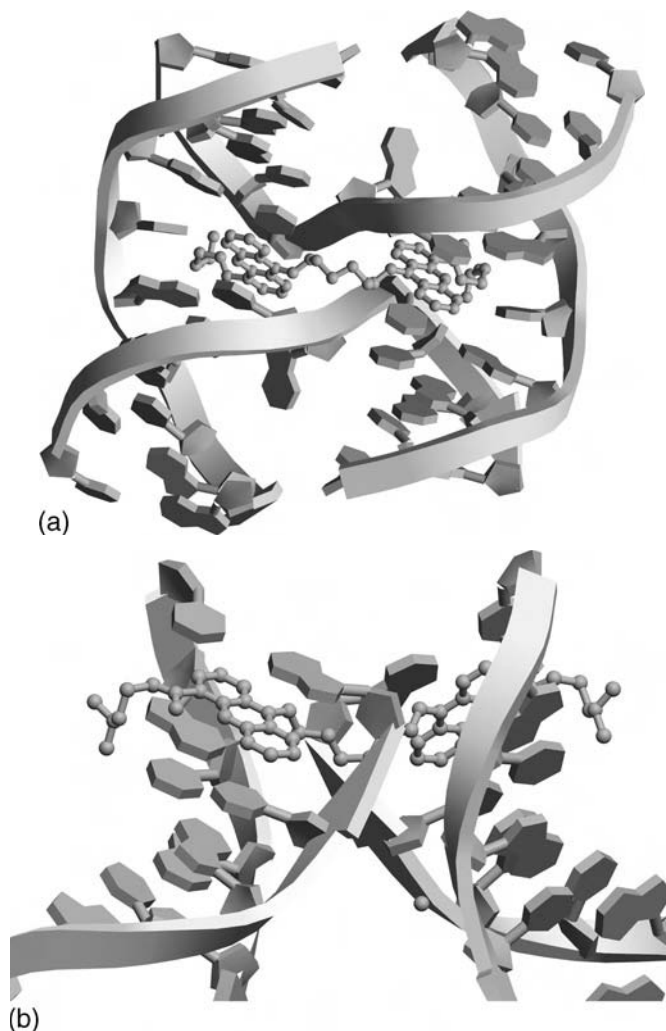


Figure 5.37 Two views of the crystal structure (Brogden *et al.*, 2007) of a complex between a bis-acridine molecule and a Holliday junction. (a) shows an overall view, and (b) shows the detailed environment around one of the acridine chromophores, which has displaced an adenine base, creating a cavity for itself.

the binding site was created by the movement of two symmetry-related adenine bases out of their stacked positions, creating a cavity for the drug. The challenge posed by this structure is whether selective molecules for junction structures can be devised from it, and if so, whether such ligands can have useful biological or therapeutic functions.

Holliday junctions are frequently used as the defining link motif in artificial supramolecular DNA assemblies (Seeman, 2005), although as yet ligand-directed assembly has not been seriously explored. The unexpected finding that four-way Holliday junctions can be formed by far shorter sequences than was hitherto envisaged, has led not only to study of the ligand complexes outlined above, but also to the realization that under some circumstances, ligand shape itself can control what type of DNA arrangement is formed. A graphic illustration of this is the formation of a three-way DNA junction by an iron-containing molecule consisting of three bis-pyridylimine groups coordinated to two Fe^{2+} ions. The crystal structure (Oleksi *et al.*, 2006) of this complex bound to the simple hexanucleotide sequence d(CGTACG) shows that a three-way DNA junction has formed around it (Fig. 5.38). Each arm contains two G•C and one A•T base pair; each of these form extensive π - π stacks with the phenyl rings of the ligand. The features of the junction, such as B-DNA arms inclined at an angle of 110° , are closely similar to those observed in the crystal structure of the Cre recombinase complex (Woods *et al.*, 2001).



Figure 5.38 The crystal structure (Oleksi *et al.*, 2006) of a three-way junction involving, at the center of the structure a ligand formed by three bis-pyridylimine groups coordinated to two Fe^{2+} ions (shown as shaded spheres).

5.6 Groove-Binding Molecules

A large and chemically diverse family of compounds can be classified as DNA groove-binders. Many show biological activity, and several of them find medicinal use as antiparasitic or antiviral agents. They generally show a preference for binding to A/T regions of DNA (Zimmer and Wähnert, 1986; Dervan, 1986). By contrast with intercalating drugs, they do not significantly perturb DNA structure. They bind exclusively in the minor groove of B-DNA duplexes. They can function as simple blockers of transcription, or as inhibitors of DNA topoisomerase enzymes. This alone is probably insufficient to explain why some minor-groove binders work as drugs. It has been suggested that for those drugs with high therapeutic indices in diseases where an external organism is the causative agent (e.g. in microbial infections), preferential drug-binding would occur in extended regions of A/T sequence in the organism. Such regions have been found in the mitochondrial DNA of these organisms and the drugs may therefore be selectively inhibiting their electron transport functions. Groove-binding ligands are also attracting increasing interest as starting points for the design of sequence-specific molecules capable of the recognition of unique sequences within a genome (see Sect. 5.1). This would have general applicability to a wide range of human diseases.

A notable exception to the A/T sequence selectivity of minor-groove drugs is illustrated by the anticancer antibiotic chromomycin, comprising five sugar rings linked to a tricyclic, mostly aromatic chromophore; unsurprisingly early biophysical data on DNA-binding was interpreted as showing intercalative-binding. The drug has high selectivity for a binding site containing at least three G•C base pairs, which is not readily explicable by such a model. The structural basis for the sequence selectivity has been revealed by several structural studies, the most recent of which is a cocrystal structure (Hou *et al.*, 2004) of a complex with the sequence d(TTGGCCAA). This has confirmed and extended the earlier NMR studies (Gao, Mirau, and Patel, 1992), and also a more recent analysis (Gochin, 2000) of a complex with the same DNA sequence as in the crystal structure. The crystal structure (Fig. 5.39) shows that chromomycin binds to the widened G/C minor groove of the B-type duplex as a dimer. A magnesium ion plays a crucial role by interacting with the two chromomycin chromophores; the two magnesium-coordinated water molecules also hydrogen-bond with various cytosine oxygen atoms in the minor groove. The G/C specificity of chromomycin is a consequence of a number of hydrogen bonds between oxygen atoms at the chromophore edge and N2 substituent atoms of guanine bases at the floor of the minor groove. The DNA, especially the recognition site itself, is appreciably distorted from B-form ideality, with a pronounced kink, and ca 10° helical unwinding.

5.6.1 Simple Groove Binding Molecules

Molecules which have been shown to bind preferential to A/T-rich duplex DNA, but not to duplex RNA or A-form DNA include synthetic DNA stains typified by Hoechst 33258 and DAPI, antitrypanocidal agents such as berenil and the drug pentamidine (Fig. 5.40). Biophysical and footprinting studies have demonstrated that these molecules bind with approximately the same affinity to DNA as

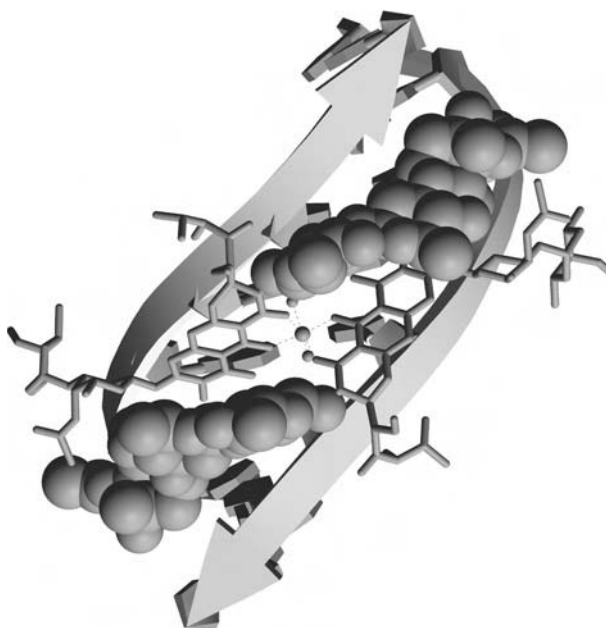


Figure 5.39 Crystal structure (Hou *et al.*, 2004) of a complex between chromomycin and an octanucleotide duplex. The trisaccharide moiety of the drug is bound in the minor groove and shown in van der Waals radius representation. The magnesium ion that coordinates to two chromomycin molecules is shown at the center of the structure.

intercalators (with typical binding affinities of 10^6 mole⁻¹), but do not perturb DNA structure.

The common structural characteristics shared by most minor-groove binding molecules are:

- Positive charge(s)
- Linked rather than fused aromatic and/or heteroaromatic rings
- An approximately crescent shape in three dimensions

A large number of X-ray crystallographic and NMR studies have shown that these molecules bind into the minor groove of B-type DNA duplexes, and that these common features play an important role in the interactions (see, e.g., Kopka *et al.*, 1985; Patel and Shapiro, 1986; Lane *et al.*, 1991; Brown *et al.*, 1992).

The majority of these structures involve complexes with duplexes formed by dodecanucleotides, most often with the Dickerson–Drew sequence d(CGCGAATTCGCG) and closely related ones such as d(CGCAAATTTGCG). Typically, the ligand is bound in the A/T region, with the aromatic groups of the ligand lying between and parallel to the two sugar-phosphate backbones (Fig. 5.41). There is little change to the structure of the DNA upon groove binding, at least for these ligands. The minor groove tends to be exceptionally narrow in this region, and atoms forming the backbone and groove floor are in close van der Waals contact with the ligand. Much of this contact is hydrophobic in nature, and predominantly involve nonpolar atoms C1'/H1', C4'/H4', and C5'/H5' of the backbone, as seen in Fig. 5.2. NMR studies have shown that the chemical shifts induced in these protons on binding can be taken as diagnostic for groove-binding.

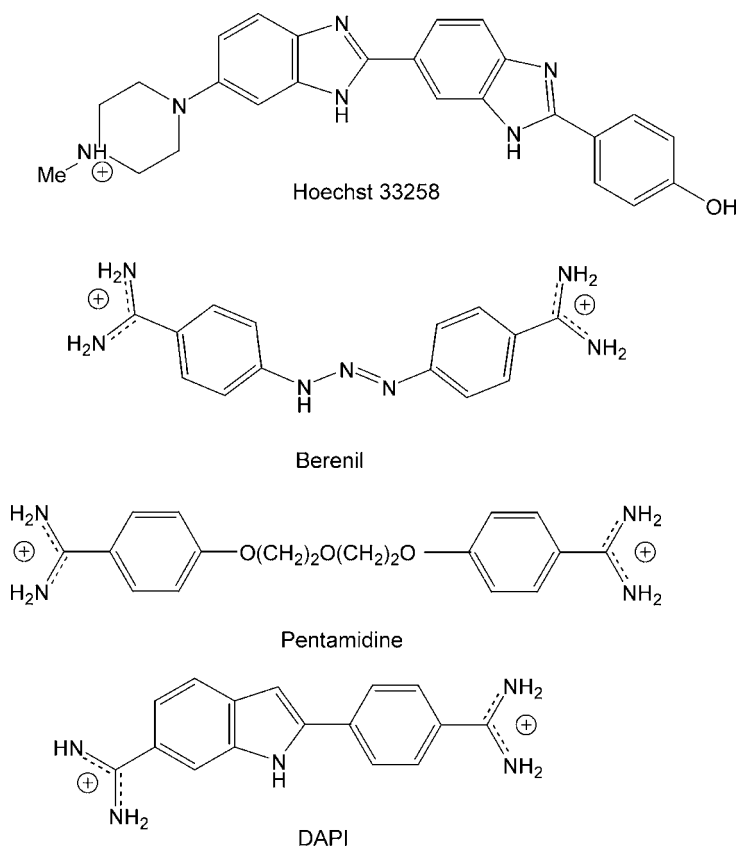


Figure 5.40 Some typical groove-binding molecules.

Hydrogen-bonding between ligand and DNA is frequently observed, notably between a donor atom on a ligand hetero ring or one in a charged terminal amidinium group (Fig. 5.42). Such hydrogen-bonding is to a thymine O2 or an adenine N3 atom. This is termed direct sequence readout, and is highly directional. Conversely, sequence readout can also be indirect, with the ligand being able to recognize particular backbone or other sequence-dependent structural features. For A/T sequences indirect readout is a consequence of (i) the narrow cross section of groove-binding molecules complementing groove width and (ii) the negative electrostatic potential in the AT minor groove complementing the positive charge(s) on these molecules. Groove interaction also involves at least part of the inner surface of the bound molecule in close contact with the convex surface of the floor of the DNA minor groove itself. Thus a G/C sequence with an exocyclic NH₂ amino group of guanine protruding into the groove will hinder the effective binding of a molecule with a smooth concave inner surface. The observed matching of ligand curvature to that of an A/T minor-groove surface for many minor-groove binding molecules has been termed “isohelicity” (Goodsell and Dickerson, 1986), and involves the shape of the drug inner surface complementing the twists in the curvature of the floor of the minor groove that are a consequence of the helical nature of the DNA double



Figure 5.41 A view of the crystal structure (Spink *et al.*, 1994) of the complex between Hoechst 33258 and the duplex formed by d(CGCAAATTTGCG), with the atoms of the drug molecule shown shaded.

helix. Isohelicity (Fig. 5.43) has been found to be a useful concept in the design of many novel groove-binding agents, although recent studies (see below) have shown that it is by no means universal.

There has been controversy concerning the relative importance of the various factors contributing to sequence-selective minor-groove binding. A number of ligands have been shown to have minimal or even no hydrogen-bonding to the edges of base pairs, and electrostatic factors are clearly of lesser importance in those instances where the ligands are uncharged. Van der Waals and hydrophobic interactions with groove walls are probably the dominant factors in overall stability when ligand is bound in the groove, and is driven in part by groove-width structure and flexibility. Directed hydrogen-bonding, though a smaller contributor to overall ligand stabilization in its binding site, is responsible for ensuring binding to particular sites and sequences within an overall sequence type. Conversely, ligands that have a general A/T selectivity and do not discriminate between different types of A/T sequence, are not involved in significant hydrogen-bonding to bases. Hydrogen-bonding is a crucial factor that can be exploited in the design of molecules with altered sequence recognition capabilities (see below).

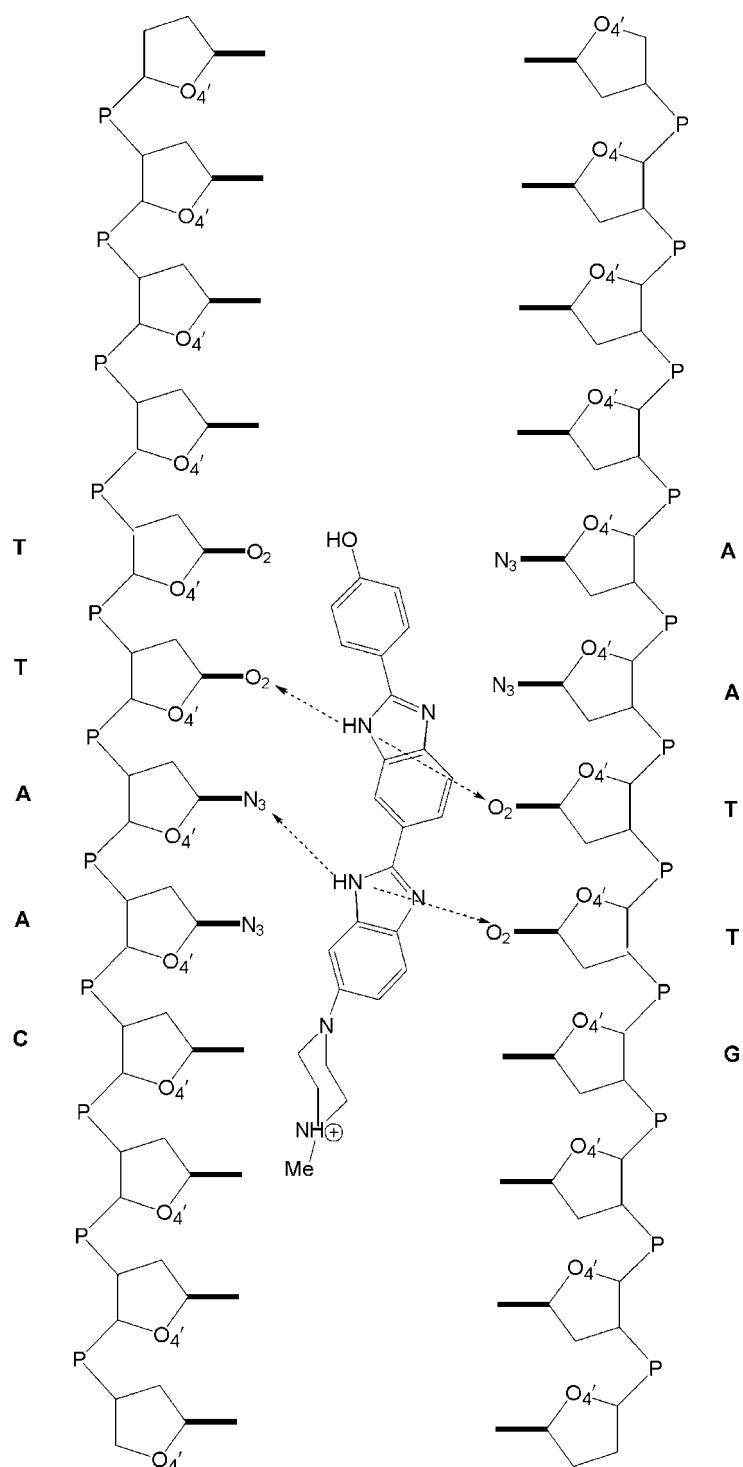


Figure 5.42 Schematic view of the hydrogen bonding between base edges and the benzimidazole groups, in the crystal structures of the Hoechst 33258 molecule complexed with duplex sequences containing the sequenceAATT....

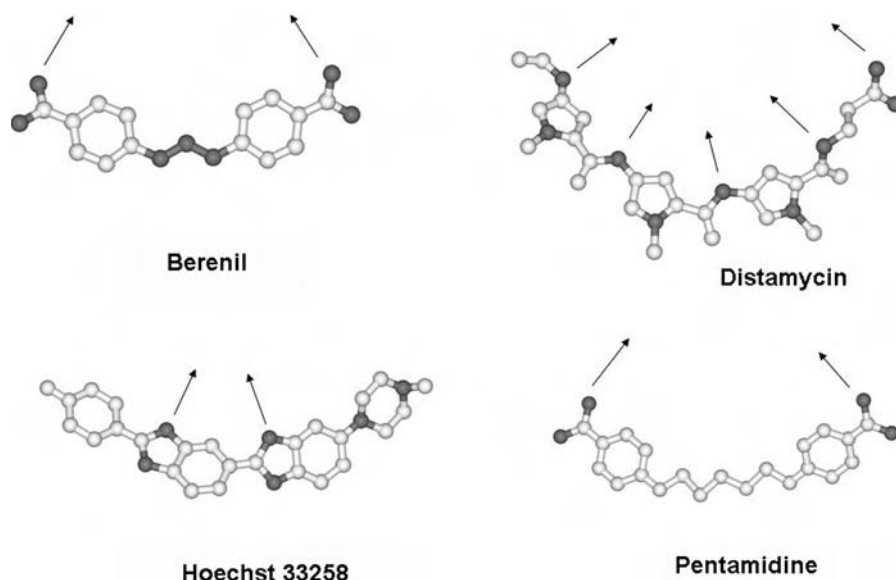


Figure 5.43 The structures of four minor-groove drugs, taken from the crystal structures of their oligonucleotide complexes (see Table 5.5). In each case their concave surface is apparent; the degree of concavity differing from one drug to the other, with distamycin having the greatest curvature. Their potential for hydrogen-bonding is indicated by an arrow for each donor – note that the maximum number of hydrogen bonds is not always used.

Another factor in ligand-binding to be considered at A/T sequences is the displacement of the highly structured arrangement of water molecules, the “spine of hydration.” The energetic driving force for this displacement is the dominance of the nonpolar interactions between ligand and groove walls. Variations in groove width are probably responsible for the greater affinity of molecules such as berenil, netropsin, Hoechst 33258, for sequences containing 5'-AATT compared to 5'-TTAA or 5'-TATA (Laughton and Luisi, 1998; Abu-daya, Brown, and Fox, 1995).

A well-studied subclass of groove binders comprises molecules with two charged amidinium groups, one at each end. Berenil is typical, with a triazine group linking two phenyl amidinium moieties. Crystallographic and NMR analyses have shown that the molecule covers 3–4 A•T base pairs (Lane *et al.*, 1991; Brown *et al.*, 1992). The former also find that each amidinium group has a N–H bond facing into the groove, which hydrogen-bonds with thymine O2 or adenine N3 base-edge atoms. In the complex with the Dickerson–Drew sequence, a water molecule mediates this interaction at one end of the binding site. A number of other bis-amidinium drugs, notably those with a flexible linker such as the drug pentamidine, bind in a similar manner. Pentamidine is of considerable clinical importance, since it has activity against the *Pneumocystis carinii* pathogen, which is responsible for the opportunistic and life-threatening pneumonia that affects the majority of AIDS patients. The drug probably works by selectively inhibiting the topoisomerases of the pathogen, via binding to its A/T-rich genome at points of topoisomerase selectivity. Although pentamidine is one of the drugs of choice, its effectiveness is limited and it produces a number of toxic side effects. Accordingly, there have been numerous studies directed to finding more selective and effective analogues, for which there are good correlations between DNA-

binding affinity and biological efficacy. Dicationic diarylfuran molecules (which are almost isostructural with berenil) have shown particular promise in this regard, and the cyclohexyl derivative (Fig. 5.44) is 100-fold more active than pentamidine itself. A crystallographic study of an oligonucleotide complex with this compound has provided a rationale for its improved DNA-binding activity, and indirectly for its biological superiority (Boykin *et al.*, 1995). The bulky cyclohexyl groups fit snugly into the minor groove, and make extensive contacts with the nonpolar atoms lining the groove walls, to a greater extent than analogues with smaller attached groups.

One particular compound in the substituted diarylfuran series, DB289, a prodrug of 2,5-bis(4-amidinophenyl)furan (furamidine), is highly active against trypanosomes, and is currently in phase II clinical trials (Wilson *et al.*, 2005). It is believed to exert its action by selectively binding to A/T-rich kinetoplast DNA in trypanosomes; the crystal structure of furamidine bound to the Dickerson–Drew dodecamer (Laughton *et al.*, 1996), shows that it forms a conventional isohelical complex. On the other hand, a number of ligands have subsequently been discovered with high A/T affinity (and often corresponding good biological activity against parasitic targets) yet which do not obey the isohelicity principle since they have a linear or near-linear shape. One example is the almost linear meta-substituted diamidine molecule CGP40215A (Fig. 5.45). The crystal structure of a complex with d(CGCGAATTCGCG)₂ shows

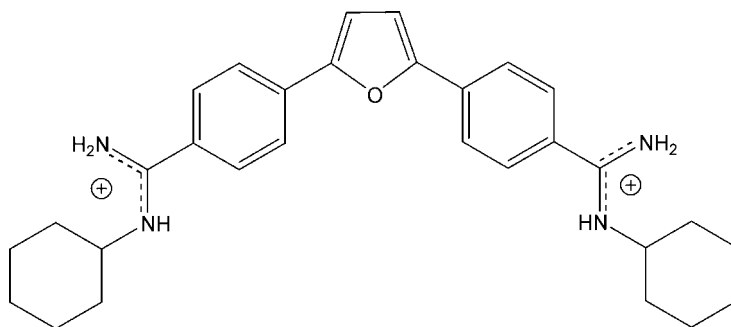


Figure 5.44 The structure of the cyclohexyl derivative of bis-[amidino-phenyl] furan (Boykin *et al.*, 1995).

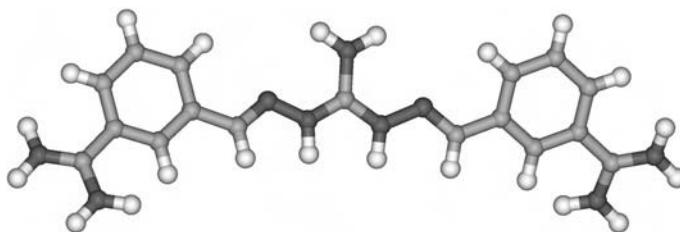


Figure 5.45 The molecular structure of CGP40215A, with nitrogen atoms shown in dark shade. Its near-linearity apparent when compared to isohelical groove-binding molecules such as berenil.

that this ligand binds in the minor groove A/T region (Nguyen *et al.*, 2002, 2004), in accord with biophysical data showing that it also has high A/T affinity and selectivity. The two central NH groups of the ligand form strong hydrogen bonds to O2 atoms of thymine bases, but the ends of the ligand only make one direct DNA contact in total. This end also makes a short water-mediated contact with base edges while the other end of the ligand is disordered and makes no DNA contacts at all (Fig. 5.46). Complementary molecular dynamics calculations are in accord with the crystal structure, and also indicate that the ligand binds in a seesaw-like manner, pivoted around the central set of ligand–DNA hydrogen bonds.

The linear ligand DB921 has a twisted biphenyl group linked in a linear way to a benzimidazole ring, giving it a noncurved structure that does not fit the isohelicity concept, yet it also binds effectively to A/T DNA. The crystal structure of its DNA complex (Miao *et al.*, 2005) shows that a water molecule again mediates between one end of the ligand and a DNA base edge (Fig. 5.47), thus effectively creating

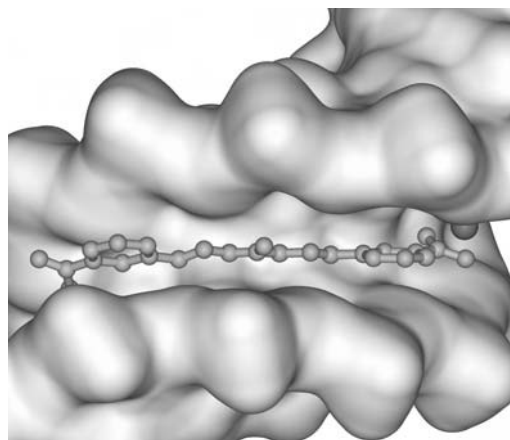


Figure 5.46 The structure of a complex of CGP40215A (shown in ball-and-stick representation) with d(CGCGAATTCGCG)₂, showing the solvent accessible surface of the DNA minor groove with the ligand bound in the A/T region. A bridging water molecule is shown buried in the groove, as a shaded sphere (Nguyen *et al.*, 2002, 2004).

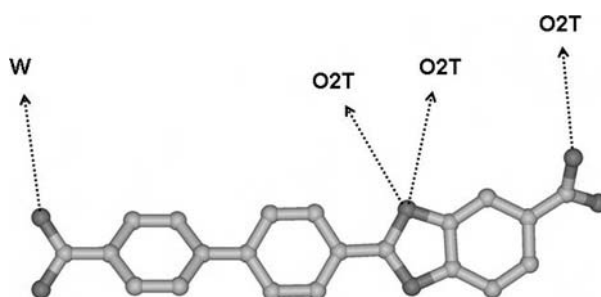


Figure 5.47 Schematic showing the hydrogen-bonding to base edges from the linear ligand DB921, as found in the crystal structure of its DNA complex (Miao *et al.*, 2005).

flexible and mobile isohelicity for this otherwise linear drug molecule. How these water molecules relate to the spine of hydration has yet to be fully explored, but it is clear that they play a more general role in A/T sequence-selective recognition (Bailly *et al.*, 2003). It is plausible to think of those that are observed in various crystal structures represent the residue left after ligand displaces the bound water network.

The bis-benzimidazole compound Hoechst 33258 is widely used as a DNA and chromosomal stain. Footprinting studies have shown that its 4–5 base-pair binding site, although necessarily mostly A/T-containing, does have a G/C pair at one end. Structural studies (Vega *et al.*, 1994; Spink *et al.*, 1994; Bostock-Smith, Laughton, and Searle, 1999; Gavathiotis, Sharman, and Searle, 2000) have provided an explanation; the two benzimidazole groups in the molecule form a network of bifurcated hydrogen bonds to three consecutive A•T base pairs (Figs. 5.41, 5.42). However the piperidine ring is nonplanar and is too bulky to fit into the narrow A/T groove region. Instead, it forces the molecule as a whole into a site with a widened groove at the 3' end, which is best accommodated by a G•C base pair. The head-to-tail arrangement of the two benzimidazole units in Hoechst 33258 has been extended to three such units, as in the molecule shown in Fig. 5.48. This extended ligand (with the trivial name TRIBIZ) binds tightly to a site of 7.5 base pairs (Fig. 5.49). As in the parent Hoechst 22358, each benzimidazole group hydrogen-bonds to two consecutive A•T base pairs by means of a pair of bifurcated hydrogen bonds. The TRIBIZ molecule is just about able to maintain all base pairs in hydrogen-bonding register, with some changes in local DNA structure being needed to achieve this (Clark *et al.*, 1996; Ji *et al.*, 2001). Further benzimidazole units linked in the same manner would not maintain the correct phasing of hydrogen bonds to successive DNA base pairs. This is a general problem, of “keeping in register” with successive base pairs in a DNA sequence. It is especially significant when designing ligands to recognize upward of a complete turn of double helix, since even small differences in DNA structure and flexibility will be magnified over longer lengths of DNA.

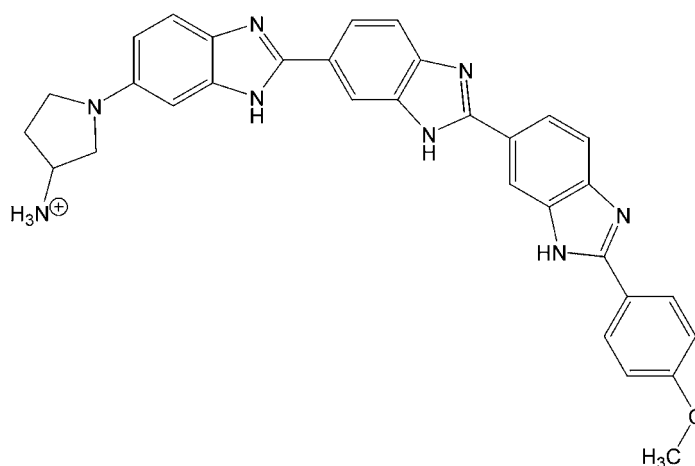


Figure 5.48 The structure of the groove-binding molecule TRIBIZ, containing three linked benzimidazole units (Clark *et al.*, 1996).



Figure 5.49 A view of the crystal structure (Clark *et al.*, 1996) of the complex between the TRIBIZ molecule and the duplex formed by d(CGCAAATTTGCG). The methoxyphenyl group of TRIBIZ is at the upper end of the binding site in this view.

5.6.2 Netropsin and Distamycin

These two classic groove-binding naturally occurring antibiotics (Fig. 5.50) have been extensively characterized, with their DNA interactions studied by a wide range of biophysical methods (Zimmer and Wähnert, 1986). Both comprise linked *N*-methyl pyrrole and amide units, with netropsin having two cationic charges compared to the one of the larger distamycin molecule. Crystal structures of both drugs complexed with A/T-containing oligonucleotides show that the narrow cross section of both complements the narrow cross section of the minor groove, again providing an explanation for their observed A/T preferences. Each amide group has its nitrogen atom oriented into the groove and these participate in (often bifurcated) hydrogen-bonding with donor atoms on the edges of consecutive adenine and/or thymine bases (Kopka *et al.*, 1985). This has been termed Class I binding. Class II binding, which has only been rarely observed, involves disordered drug ends and a consequent reduction in sequence selectivity. A number of subsequent crystallographic studies on netropsin complexes (see, e.g., Nunn, Garman, and Neidle, 1997; Van Hecke *et al.*, 2005), have confirmed the Class I structural features, although there are sometimes small differences of detail depending on the precise A/T target sequence and also on the nature of the flanking sequences. A recent approach has used the host-guest technique outlined in Chap. 4, in which the target oligonucleotide is effectively pinned between two host protein

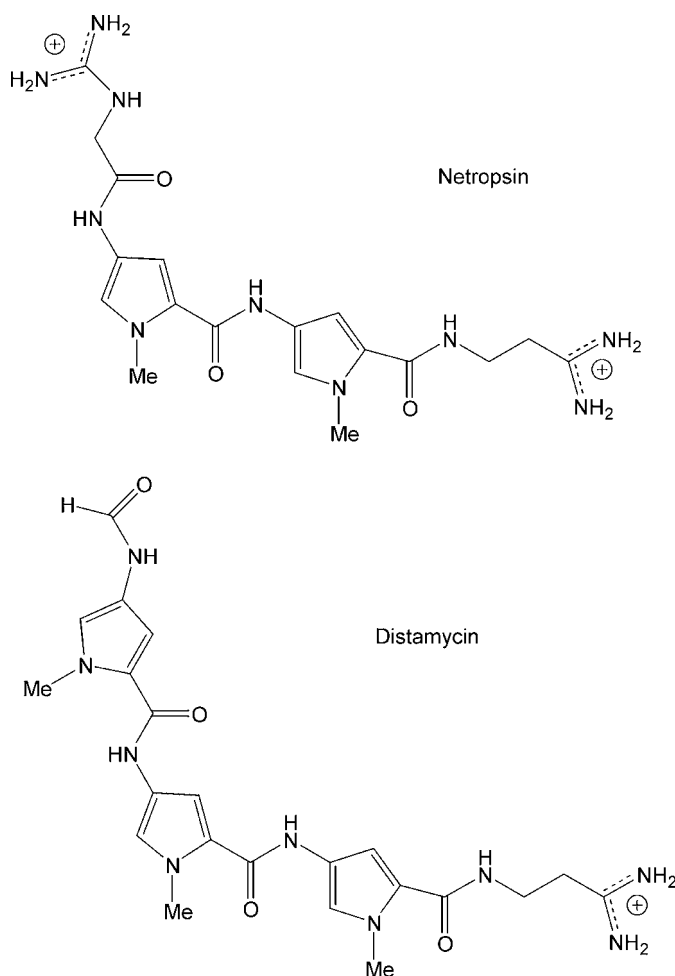


Figure 5.50 The structures of the oligopeptide-like groove binding drugs netropsin and distamycin.

molecules (Goodwin, Long, and Georgiadis, 2005). The crystal structure of such a complex involving netropsin and the oligonucleotide d(CTTAATTCGAATTAAG), illustrates the advantage of this approach, showing the two AATT drug binding sites in the one sequence (Fig. 5.51). Isomorphous structures with variants of the target sequence and changes in ligand type can be readily obtained; the method has also been applied to a diamidine derivative of benzimidazole (Goodwin *et al.*, 2006; Tanious *et al.*, 2007).

All of these groove binders show a consistent pattern of interaction with A/T stretches of duplex DNA, with two principal sets of interactions dominating the sequence selectivity: shape complementarity with the narrow groove walls, and either direct or water-mediated hydrogen-bonding to base edges at the groove floor. Implied also is a negative factor, that the absence of guanine exocyclic -NH_2 substituents on the floor of the groove ensures that the shape and hydrogen-bonding complementarity are optimal at A/T sites. This concept has been used in the design of molecules with potential G/C selectivity at particular points (Fig. 5.52).

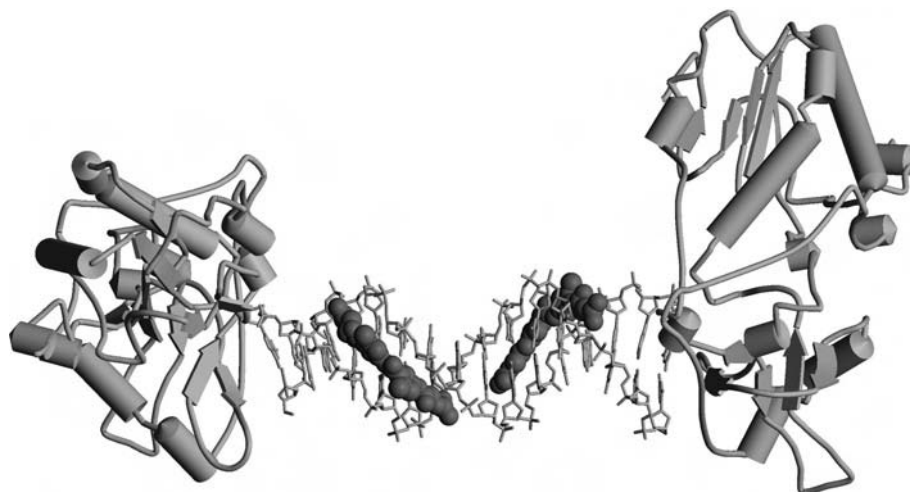


Figure 5.51 The crystal structure (Goodwin, Long, and Georgiadis, 2005) of a host-guest complex involving netropsin (in space-filling mode) and the oligonucleotide d(CTTAATTTCGAATTAAAG), showing the two AATT drug-binding sites in the one sequence. The protein host molecules at the end of the duplex are also shown, in cartoon form.

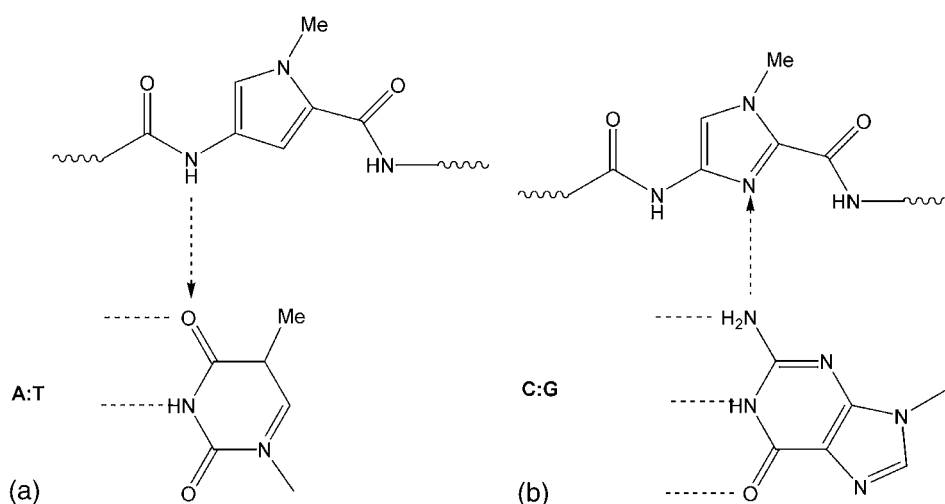


Figure 5.52 (a) The principles of netropsin and distamycin amide-base recognition, showing hydrogen-bonding to the thymine of an A•T base pair. (b) The concept of lexitropsin base recognition, with hydrogen-bonding from the exocyclic amino substituent of a guanine to the nitrogen atom of an imidazole ring.

Such molecules, termed “lexitropsins,” are based on the proposal (Lown, 1994) of switching hydrogen-bond polarity at the groove floor by means of, for example, an imidazole ring in order to hydrogen-bond to the guanine NH_2 substituent. In practice, lexitropsins have only rarely completely achieved the goal of a total change in sequence specificity. Instead most bind to both A/T and mixed sequences (Kopka *et al.*, 1997), and tend to show reduced affinities for both types of site. In retrospect this failure can be ascribed to the differences in both groove width and flexibility in

G/C compared to A/T regions. The latter tend to be wider and a typical groove-binding molecule with a narrow cross section will not bind well. The astute reader will also notice that the idealized hydrogen-bonding scheme for lexitropsins implies subtle differences in position for hydrogen-bond donors and acceptors, which are difficult to take into account in the design process. In spite of these problems, the lexitropsin concept has proved its value when combined with dimeric groove-binding, as discussed below.

An important observation, initially from NMR studies (Pelton and Wemmer, 1990), was that at higher drug:DNA ratios, distamycin can form a 2:1 complex with the duplex sequence d(CGCAAATTTGCG). This has been of major significance for the subsequent design of the polyamide class of sequence-selective ligand (Table 5.5). This NMR-derived structure has been characterized in detail, and crystallographic analyses have been reported for a number of other 2:1 complexes with distamycin bound to a range of sequences (e.g., Chen *et al.*, 1994). All show the drug bound in the minor groove as a dimer, in an antiparallel head-to-tail manner (Figs. 5.53, 5.54), in striking contrast to the pattern of 1:1 complexes described above, and in particular to the structure of the 1:1 distamycin-d(CGCAAATTTGCG) complex. The positively charged ends of the two distamycin molecules are far apart in the dimer complexes, suggesting that this arrangement could occur with other groove-binding molecules possessing a single positive charge, such as Hoechst 33258; as yet no such observations have been reported. The NMR and crystallographic structures of the 2:1 complexes have both distamycin molecules involved in close nonbonded contacts with each other, and most importantly, each interacts with just one DNA strand (Fig. 5.55). In the 1:1 complex of distamycin (and also those of netropsin and Hoechst 33258), hydrogen-bonding to base edges involves sets of three-center bifurcated hydrogen bonds between an amide group and any two of adenine N3 and thymine O2 atoms. In order for both distamycin molecules to be accommodated in the 2:1

Table 5.5 Crystal Structures of Drug-Oligonucleotide Minor-Groove Complexes

Drug	Sequence	PDB ID Code	NDB ID code
Distamycin	d(CGCAAATTTGCG)	2DND	GDL003
Netropsin	d(CGCGAATTCGCG)	6BNA	GDLB05
Hoechst 33258	d(CGCGAATTCGCG)	1D43, 1D44	GDL010, GDL011
Netropsin	d(CTTAATTCGAATTAAG)	1ZTT	PD0678
Hoechst 33258	d(CGCAAATTTGCG)	264D, 296D	GDL026, GDL028
Berenil	d(CGCGAATTCGCG)	2DBE	GDL009
Pentamidine	d(CGCGAATTCGCG)	1D64	GDL015
Furamidine	d(CGCGAATTCGCG)	227D	GDL036
Cyclohexyl-furamidine	d(CGCGAATTCGCG)	1FMS	DD0035
CGP40215A	d(CGCGAATTCGCG)	1M6F	DD0052
DB921	d(CGCGAATTCGCG)	2B0K	DD0074
Meta-OH-Hoechst	d(CGCGAATTCGCG)	302D, 303D	GDL047, GDL048
TRIBIZ	d(CGCAAATTTGCG)	263D	GDL039
DAPI	d(CGCGAATTCGCG)	1D30	GDL008
Mono-Im lexitropsin	d(CGCGAATTCGCG))	1LEX, 1LEY	GLD037, GLD038
Di-imidazole lexitropsin	d(CATGGCCATG)	334D	GDJ054
2:1 Distamycin	d(ICICICIC)	159D	GDHB25
Polyamide Im-Hp-Py-Py	d(CCAGTACTGG)	407D	BDD002
Polyamide Im-Py-Hp-Py	d(CCAGATCTGG)	1CVX	DD0020
Chromomycin	d(TTGGCCAA)	1VAQ	DD0063

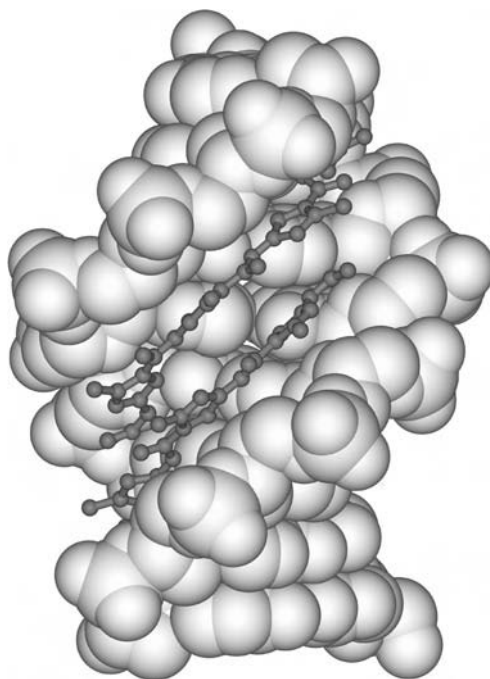


Figure 5.53 The crystal structure (Chen *et al.*, 1994) of a 2:1 complex between distamycin (with atoms colored black) and the duplex formed by d(ICICICIC).

complex, the “narrow minor groove” involving the 5′-AAATTT tract must expand from 3.4 Å to 6.8 Å. Thus, a new picture emerges, so that rather than the minor groove having rigid geometry, we have to consider it as being inherently flexible, especially in order to accommodate ligand-binding (Bostock-Smith, Laughton, and Searle, 1999).

5.6.3 Sequence-Specific Polyamides

The 2:1 mode of distamycin-binding is the basis for a large number of subsequent studies which have made attainable the goal of sequence recognition at the gene level (Mrksich and Dervan, 1993; Dervan and Bürli, 1999). The key feature demonstrated by the 2:1 complexes is that minor-groove recognition can involve simultaneous hydrogen-bonding to *both* bases in a base pair. Thus far greater discrimination is inherently attainable than can be achieved with 1:1 minor-groove binders, and more than might be envisaged by solely considering canonical B-DNA minor-groove width, which in the absence of a ligand that can widen it, seems incapable of binding a ligand large enough to recognize both bases in a pair. A general approach for switching from A•T to G•C base-pair recognition has been to adopt a lexitropsin-type of change, whereby one or more *N*-methyl pyrrole (Py) groups is changed to an imidazole (Im) one (Figs. 5.52b, 5.55).

A general recognition code has been established from these studies which enables a wide range of sequences to be recognized using strings of these Im and Py groups linked by amide groups (Table 5.6a).

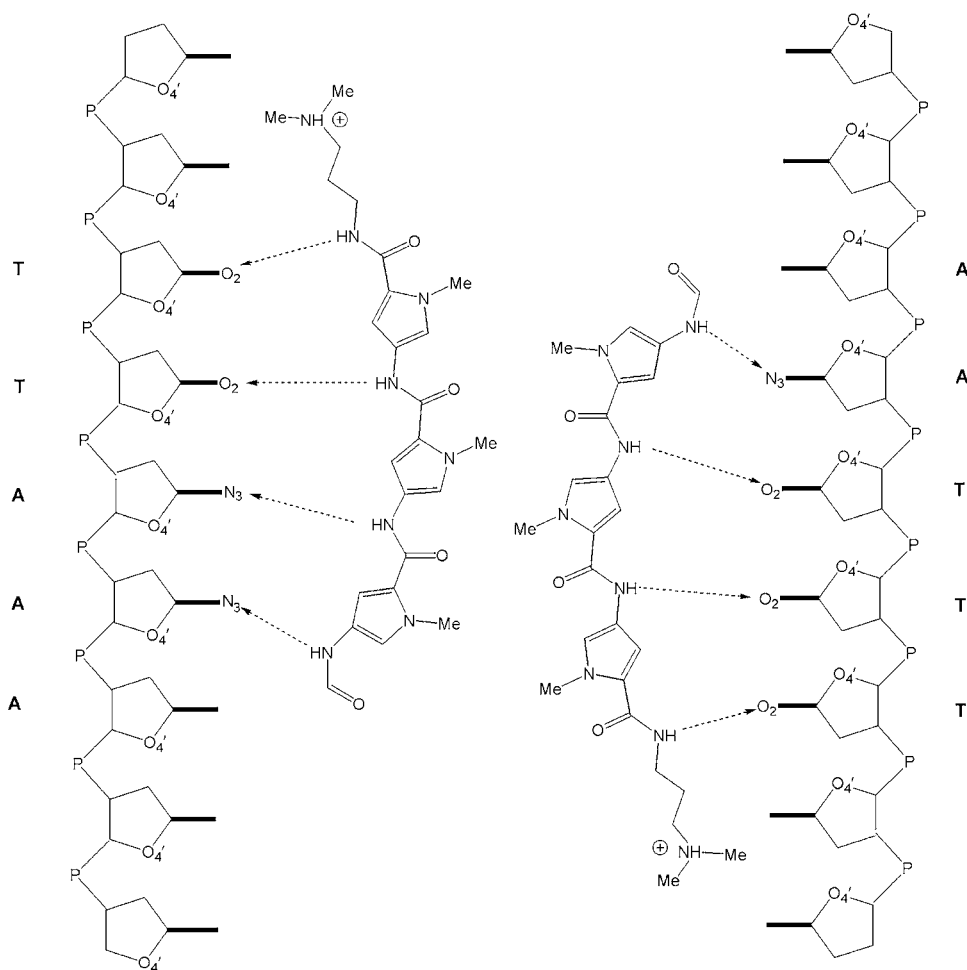


Figure 5.54 Detail of the hydrogen-bonding to both DNA strands in a 2:1 distamycin dimer showing the hydrogen-bonding to bases in both strands of a five base-pair A/T sequence.

The resulting molecules incorporating these rules in their features, termed polyamides, have been shown by footprinting analysis to be able to target, normally as dimers, a wide range of DNA sequences, and with remarkably high selectivity. Their typical specific base interactions are as shown in Fig. 5.55. Other small heterocyclic rings such as pyrrole (Py) and 3-hydroxypyrrole (Hp) have also been successfully incorporated into polyamides to provide hydrogen-bonding to particular bases, and the recognition code has more recently been extended by the use of benzimidazole (Table 5.6b) as a versatile recognition unit (Marques *et al.*, 2004). Validation of these building blocks has almost entirely been performed to date using footprinting methods; however these can only be applied to a small number of sequences at any one time. A high-throughput microarray method has recently been developed (Warren *et al.*, 2006) that is able to examine all possible sequences available to eight base-pairs of DNA, using self-complementary palindromes with a short hairpin sequence. Results from this are in excellent agreement with previous studies, emphasizing the future potential of the approach for studying small-molecule DNA sequence- and structure-specificity.

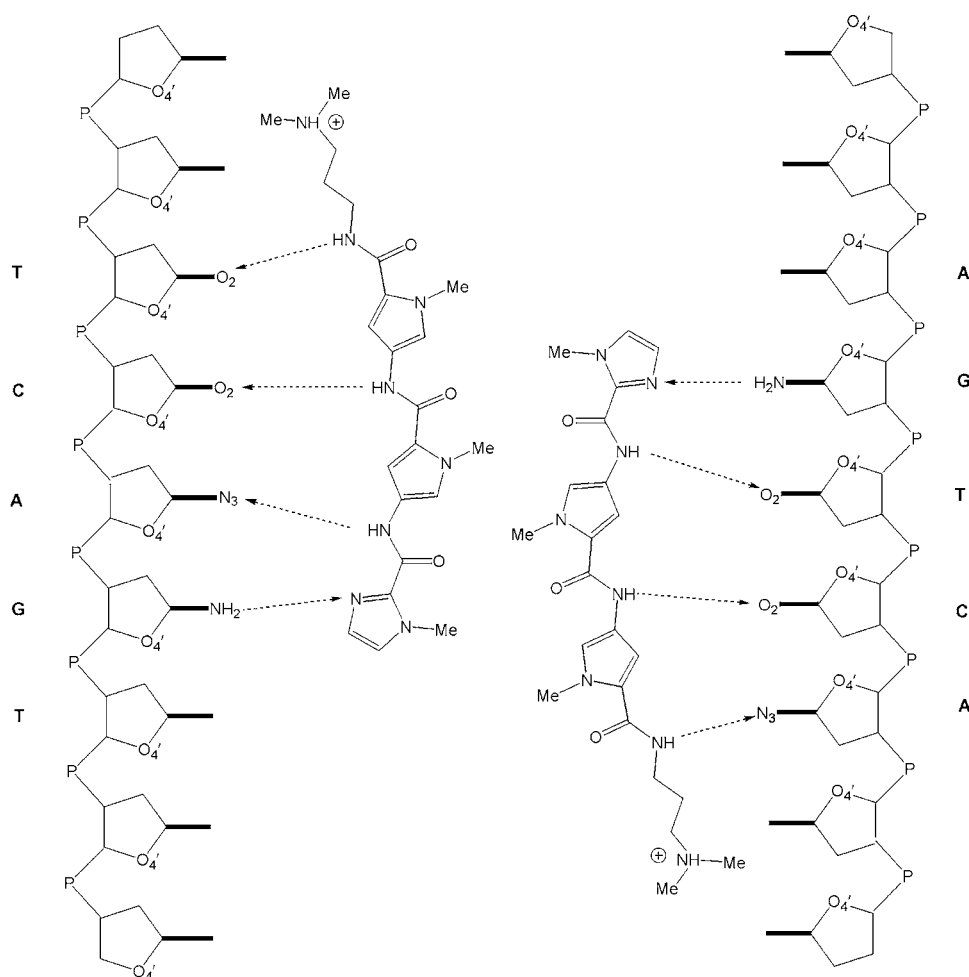


Figure 5.55 A schematic view of the hydrogen-bonding to A•T and G•C base pair edges as observed in the crystal structure (Kielkopf *et al.*, 1998) of the complex involving the polyamide Im-Py-Py and the target sequence 5'-TCAGT.

Greater control of recognition has been achieved by covalently linking two polyamide molecules, which need not have the same sequence of recognition units (Fig. 5.56). This avoids the possibility of slippage between the two unconnected components of a dimer, with potential ambiguity in sequence readout. These hairpin polyamides can achieve exceptionally high selectivity and site affinities, with typical

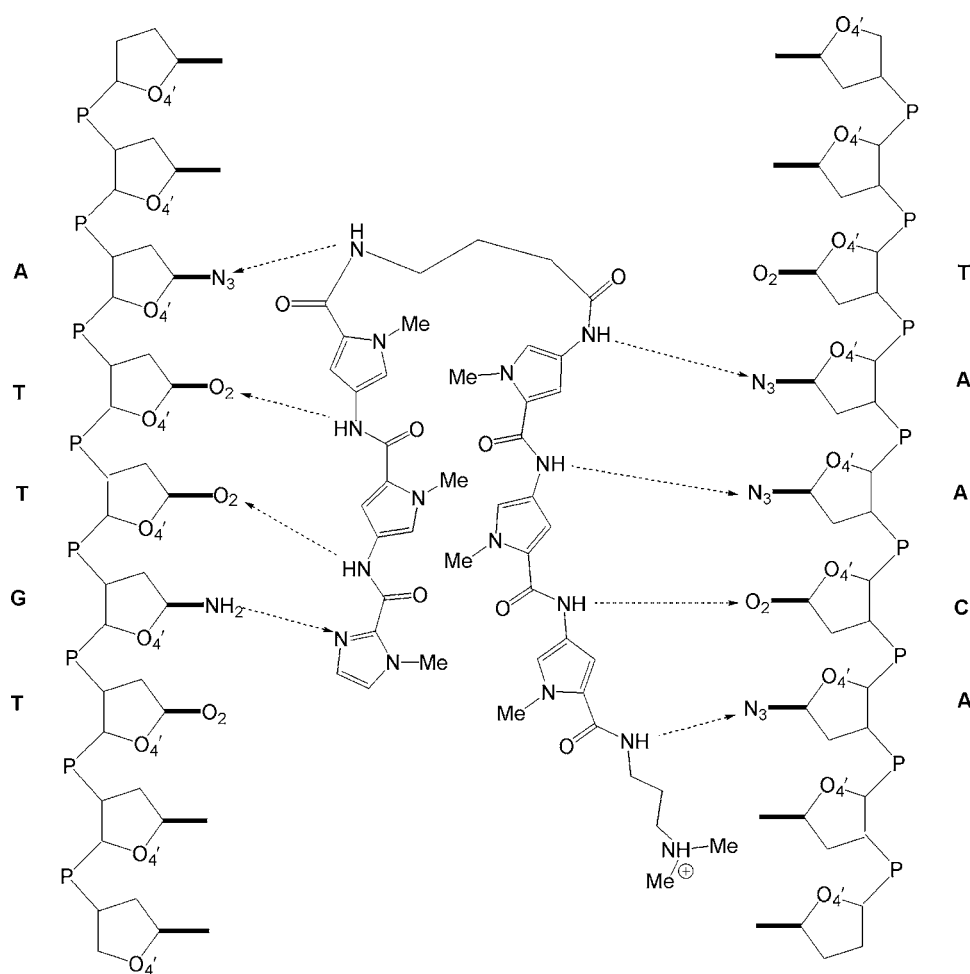
Table 5.6 (a) The Im/Py Pairing Codes for Minor-Groove Polyamide Recognition

Pair	C:G	G:C	T:A	A:T
Im/Py	–	+	–	–
Py/Im	+	–	–	–
Hp/Py	–	–	+	–
Py/Hp	–	–	–	+

Table 5.6 (b) The Benzimidazole Pairing Codes for Minor Groove Polyamide Recognition, Involving Hydroxybenzimidazole (Hz) and Imidazopyridine (Ip)

Pair	C:G	G:C	T:A	A:T
Ip/Py	–	+	–	–
Py/Ip	+	–	–	–
Hz/Py	–	–	+	–
Py/Hz	–	–	–	+

binding constants in the nM range. For example, the molecule Im-Py-Py-Py- γ -Py-Py-Py-Py- β -Dp (where γ is the hairpin linker γ -aminobutyric acid, β is β -alanine and Dp is dimethylaminopropylamide), binds to the sequence d(TGTTAT) with a dissociation constant of 1.1 nM. Several NMR and crystal structures (de Clairac *et al.*, 1997; Kielkopf *et al.*, 1998a,b, 2000; Zhang *et al.*, 2004) have been determined for

**Figure 5.56** A schematic view of a hairpin polyamide, Im-Py-Py-linker-Py-Py-Py, recognizing both strands of a G/C-containing sequence.

polyamide–oligonucleotide complexes, which have found the predicted patterns of polyamide–base recognition and thus have demonstrated the overall correctness of the approach (Fig. 5.57).

An extensive series of studies have shown that polyamides can target DNA response elements which are sites for transcription factor binding, and successfully compete with these regulatory proteins (see, e.g., Gottesfeld *et al.*, 1997, 2000; Dickinson *et al.*, 1999; Mapp *et al.*, 2000; Bremer *et al.*, 2001; Gearhart *et al.*, 2005; Burnett *et al.*, 2006; Viger and Dervan, 2006). For example, repression of the 5S RNA gene has been achieved by targeting an eight-ring polyamide to the binding site d(AGTACT) within the TFIID transcription factor binding region. This polyamide binds 30-fold more tightly than does TFIID to its complete 50-base pair sequence, and effective inhibition of transcription both *in vitro* and in cells was observed. Polyamides can also bind to specialized DNA sequences, as shown by the effects of a β -alanine-pyrrole-imidazole polyamide targeting the extensive GAA•TTC repeats found in the frataxin gene in patients suffering from the neurodegenerative disease Friedreich's ataxia. This polyamide appears to be able to lock the repeats in a B-like conformation, preventing alternative DNA structures from forming and thus aiding frataxin transcription to occur.

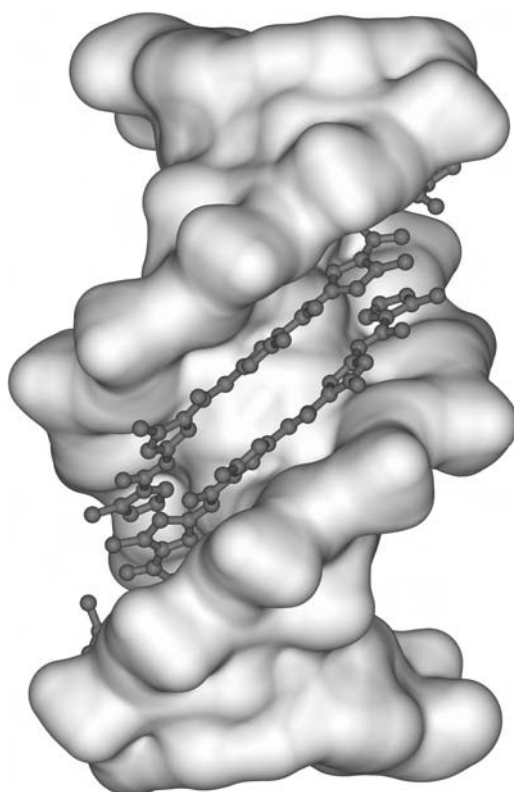


Figure 5.57 The crystal structure (Kielkopf *et al.*, 2000) of the polyamide molecule Im-Py-Py bound to a duplex formed by the sequence d(CCAGATCTGG). The central part of this sequence, i.e. 5'-AGATCT, is recognized by the polyamide.

Polyamides can also be used to upregulate transcription, by blocking repressor sites. A quite unexpected effect has been found (Henikoff and Vermaak, 2000; Janssen, Durussel, and Laemmli, 2000; Janssen *et al.*, 2000; Maeshima, Janssen, and Laemmli, 2001) with a polyamide targeted to the sequence d(GAGAAGAGAA) in the fruit fly *Drosophila*. This is a satellite DNA sequence, and is normally considered to be functionally inert. It was found that the flies had highly specific loss of function in several defined genes, which are presumed to lie close and to be affected by binding to satellite DNA. Overall these remarkable effects have been attributed to the experimental observations of drug-induced opening-up of satellite chromatin and a consequent long-range effect on specific genes. It is tempting to speculate that analogous events are involved in some of the biological effects observed for the classic A/T-binding minor groove agents and their more recent analogues.

Are polyamides going to deliver the holy grail of single gene discrimination within a genome? It is probably premature to be able to definitively answer this, since some major obstacles remain before *any* desired 16–18 base pair can be specifically recognized. For one, it is clear that the strategy does not work equally well for all sequences, and that beyond about a length of about 7–9 base pairs, the problem of maintaining polyamide subunits in register sometimes becomes severe. This may be due to differences in flexibility between DNA sequences. Most importantly for therapeutic purposes, polyamides are relatively poorly transported into some cell types compared to conventional minor-groove drugs such as Hoechst 33258 or pentamidine. A desirable goal is to devise pharmacophore-like building blocks to mimic polyamides, but as yet this work is in its infancy. The repertoire of suitable molecules has been recently extended by the finding (Wang *et al.*, 2000) that an asymmetric aromatic dication based on the diphenyl furan type of molecule (Boykin *et al.*, 1995), can also bind in the minor groove as a dimer.

5.7 Small Molecule Covalent Bonding to DNA

Historically the first DNA-interactive anticancer drugs to be developed were alkylating agents, derived from the First World War poison gases. These drugs, the nitrogen mustards, are highly toxic molecules, although their more modern analogues chlorambucil, L-phenylalanine mustard (melphalan) and cyclophosphamide still find use against some cancers. Nonspecific covalent-binding can take place on the phosphodiester backbone or sugar residues. This is often a step in DNA strand scission, as in the case of the bleomycin family of anticancer drugs. Both single- and double-strand breaks can then occur. The latter are usually lethal events to a cell since they are especially difficult to repair. There has been particular emphasis on studies of binding to particular sites on DNA bases by drugs that undertake nucleophilic reactions. Purines are the most susceptible to covalent attack, with guanine being preferred over adenine. Particularly favored sites are O6 (guanine), N6 (adenine), and N7 in the major groove and N1, N2 (guanine), and N3 in the minor groove. These site preferences are the result of the differing electronic charge distributions in bases, base pairs, and in runs of sequence. The propensity for alkylation by nitrogen mustards at N7 of a guanine base embedded within guanine-rich sequences (Hartley, Bingham, and Souhami, 1992), is an inherent property of these sequences, since the same pattern of sequence selectivity has been found both *in vitro* and in DNA extracted from mustard-treated cells.

Drugs can bind to a single site or to two at once if they have bifunctional capability. This latter cross-linking mode can be either intra- or interstrand, depending on two principal factors:

- The distance between the two functional groups on the drug
- The nature of the affected DNA sequence, for example, whether two adjacent guanine are on the same or opposite strands

Information on the structures of covalent adducts has to date been obtained from on the one hand, chemical and biophysical probe experiments, and on the other, from NMR and some X-ray crystallographic studies. Crystallization of covalent adducts has proved remarkably difficult, and significant success has mostly been achieved with platinum complexes. In accordance with the theme of this book, we shall focus here on some representative drugs where firm structural information is available, and which are of particular interest at the present time. More extensive accounts of these and the large number of other drugs in this category are available in the reading list at the end of the chapter.

5.7.1 The Platinum Drugs

The chance discovery and subsequent clinical exploitation of the cytotoxic and antitumor properties of the strikingly simple molecule *cis*-dichlorodiamminoplatinum (II) (cisplatin or *cis*-DDP), is one of the great success stories of cancer chemotherapy. It is a highly toxic drug, yet it and its closely related analogues are the probably most effective single agents in current use in the cancer clinic, where it is curative in over 90% of testicular cancers, a disease previously hard to treat, and one that had a poor prognosis.

Cisplatin mostly binds to purines, especially guanine at the N7 position. A variety of adducts are formed with DNA in solution, with the major species having intra-strand cross-links (Fig. 5.58) to the sequence d(GG), and to a lesser extent to d(AG). A number of more minor species have been characterized, but they are probably of

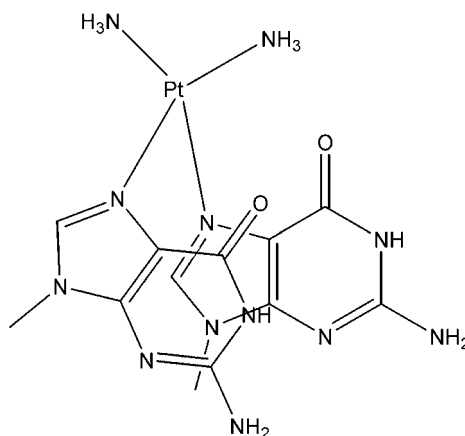


Figure 5.58 A schematic view of the covalent intrastrand cross-linking interactions of the drug *cis*-platinum with the N7 atoms of two consecutive guanine bases.

little functional significance. Several NMR and crystal structures have been determined for oligonucleotides containing cisplatin bound at d(GG) sites. The structure of a platinum-containing adduct d(CCTCTG*GTCTCC)•d(GGAGACCAGAGG), where G*G represents the adduct, has been solved by both techniques (Takahara *et al.*, 1996; Gelasco and Lippard, 1998), and provides an interesting comparison of results from solution versus the crystal. The length of the platinum-N7 bond, of ca 2.0 Å, forces the two guanines out of their normal coplanarity, and the conformations of the two adjacent guanines become highly distorted from native B-DNA. These distortions are not confined to this region of the sequence. The duplex becomes significantly bent, with a large roll toward the major groove at the *cis*-platin-binding site. The NMR study finds a roll of 49° whereas that from the X-ray analysis (Fig. 5.59) is less, of 26°. As a consequence the major groove in both structures becomes more compact and the minor groove is wider and shallower than in canonical B-DNA, changing by up to 6 Å in both width and depth. All these features are more exaggerated in the NMR structure, possibly because of the lack of crystal-packing constraints.

These large-scale structural distortions in DNA are remarkably similar to those produced in DNA by the HMG (high-mobility group) chromosomal proteins such as SRY and LEF-1, which are known to bind to the minor groove of DNA. The crystal structure has been determined for a ternary complex between a 16-base pair duplex

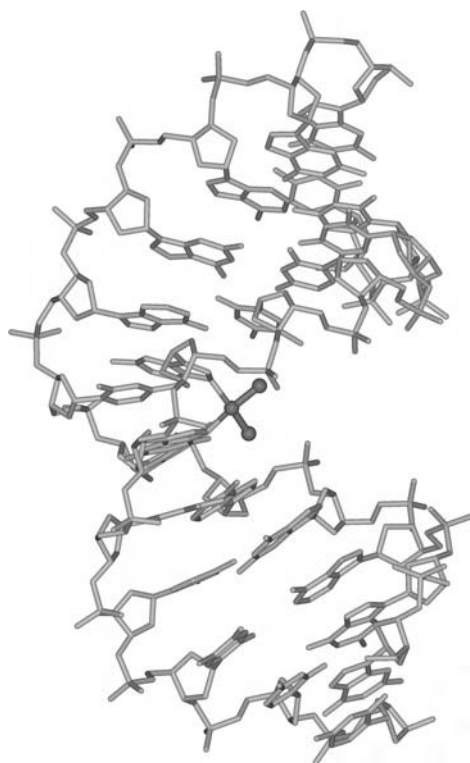


Figure 5.59 The crystal structure (Takahara, Frederick, and Lippard, 1996) of the covalent complex with cis-platinum (shown in ball-and-stick representation) and the sequences d(CCTCTG*GTCTCC) + d(GGAGACCAGAGG), where G*G represents the cross-linked adduct. Note the marked bend of the two halves of the duplex.

containing a single platinum intrastrand cross-link, and an 89-amino acid domain of the HMG1 protein (Ohndorf *et al.*, 1999). This is one of the very few drug-DNA-protein ternary complex structures to be determined. It provides a detailed picture of the interplay between the platinum-induced and protein-induced structural changes in DNA, and how platinum lesions in DNA might be recognized by repair proteins. As in the binary platinum adduct, the DNA is sharply bent toward the major groove, but this time by 61°, and the minor groove in this region resembles that in canonical A-form DNA. The protein-binding site is wholly on the 3' side of the bound platinum, and the bend itself, whereas the protein-DNA complex alone has protein bound in the center of the bend. There is more than a passing resemblance between the overall features of the structure and those of the TATA-box-DNA complex.

In striking contrast, the solution structure of DNA containing the less abundant G:G interstrand cross-link has been reported to show local unwinding at the lesion such that there is a reversal at this point to left-handedness and an extra-helical arrangement for the cytosines paired to the cross-linked guanines (Huang *et al.*, 1995). This places the platinum in the minor groove. The crystal structure (Coste *et al.*, 1999) of the same interstrand cross-link, but incorporated in the decamer d(CCTCG*CTCTC)•d(GAGAG*CGAGG), also shows these features as well as equivalent minor-groove widths, although other aspects such as the degree of unwinding at the cross-link and the degree of bending are strikingly different. This suggests that sequence context may be important. However these structural differences may also reflect the real differences between NMR and X-ray methods in their ability in being able to define long-range structural features with precision. In general, as discussed earlier, NMR methods are not as powerful in defining features such as unwinding and bending when the number of NOE distances is not large.

An even greater degree of distortion has been observed in a 1,3 G:G intrastrand cross-linked solution structure (Teuben *et al.*, 1999), with a loss of base-pairing around the lesion. Again, a pyrimidine base (here a thymine) is extruded from the helix. These consistent observations of extrahelical bases in platinum adducts suggest that they may be a general phenomenon for certain types of platinum lesion. They may act as signals for subsequent DNA repair proteins, suggesting that these minor adducts are relatively unimportant for the antitumor activity of cisplatin compared to the 1,2-intrastrand lesion, which is less readily repaired and therefore more persistent.

Continuing attempts to improve potency, reduce toxicity and resistance, and increase the range of treatable tumors has led to the development of a large number of platinum (II) and (IV) derivatives, although only two of them, carboplatin (*cis*-diammine-1,1-cyclobutanedicarboxylatoplatinum) and oxaliplatin ((1*R*,2*R*-diamino-cyclohexane) oxalatoplatinum (II)), are currently in clinical use. The crystal structure

Table 5.7 Oligonucleotide Crystal Structures with Covalently Bound Drug Molecules. In Each Case the Asterisk Indicates the Sites of Covalent Binding Along the DNA Sequence

Sequence	Drug	PDB ID code	NDB ID code
d(CCTCTG*GTCTCC) + d(GGAGACCAGAGG)	cis-platinum	1AIO	DDL873
d(CCTCG*CTCTC) + d(GAGAG*CGAGG)	cis-platinum	1A2E	DDJ075
d(CCTCTG*GTCTCC) + d(GGAGACCAGAGG)	Oxaliplatin	1IHH	DD0040
d(CCTCTG*GTCTCC) + d(GGAGACCACAGG)	Pt(ammine)- cyclohexylamine	1LU5	DD0050
d(CCAACGTTGG)	anthramycin	274D	GJJB29

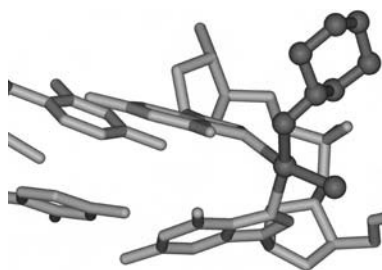


Figure 5.60 Detailed view of part of the crystal structure of a cyclohexylamine platinum (IV) adduct with a dodecamer duplex (Silverman *et al.*, 2002), showing the asymmetric 1,2-G:G intrastrand cross-link.

of a cyclohexylamine platinum (IV) adduct with a dodecamer duplex (Silverman *et al.*, 2002) shows that the drug forms an asymmetric 1,2-G:G intrastrand cross-link (Fig. 5.60). The base pairs are kept intact in spite of a large positive roll caused by the cross-linking, and overall the structure has a close resemblance to other platinum DNA complexes, including that of oxaliplatin, which also has a 1,2-G:G intrastrand cross-link (Spingler, Whittington, and Lippard, 2001). The similar pattern of DNA distortion observed in the various crystal structures is in contrast to their diverse biological activity. However a more recent NMR structure analysis has compared 1,2-G:G intrastrand cross-link adducts of cisplatin and oxaliplatin in the identical sequence (Wu *et al.*, 2007), which has revealed some differences in geometry that may be reflected in differences in recognition by repair proteins and polymerases. The base-pair steps at and close to the adduct sites show differences in a number of base morphological parameters especially helical twist and base-pair roll, that result in distinct values for the global helical bend. Although crystal-packing has been invoked to account for the differences between crystal and solution results, it is likely a consequence of distinct refinement methods combined with the fact that none of these structures approach atomic resolution.

5.7.2 Covalent-Binding Combined with Sequence-Specific Recognition

Sequence-specific DNA recognition is well illustrated in the case of the antitumor antibiotic (+)-CC-1065 (Fig. 5.61), which has three pyrrolo-indole units, and its simpler analogue duocarmycin, with just one. In both instances, one end of the drug molecule binds covalently to N3 of an adenine via opening of the cyclopropane ring (Hurley *et al.*, 1988). The rest of the drug lies in the minor groove. CC-1065 covers four base pairs to the 5' side of this adenine and one base pair on the 3' side. Specific recognition is to sequences such as 5'-AAAAA; the drug induces bending of 17–22°, comparable to that found in natural A tracts, which enables a close steric fit between drug and DNA to take place. So CC-1065 has a specific (and highly stringent) requirement for a sequence that will bend in an appropriate manner. The smaller drug molecule duocarmycin induces fewer structural changes in its target sequences, and the overall conformations remain close to canonical B-DNA (Eis *et al.*, 1997). Bizelesin is a synthetic analogue of CC-1065 with high experimental antitumor activity. It has a chemically reactive chloromethyl group at each end of the molecule, and

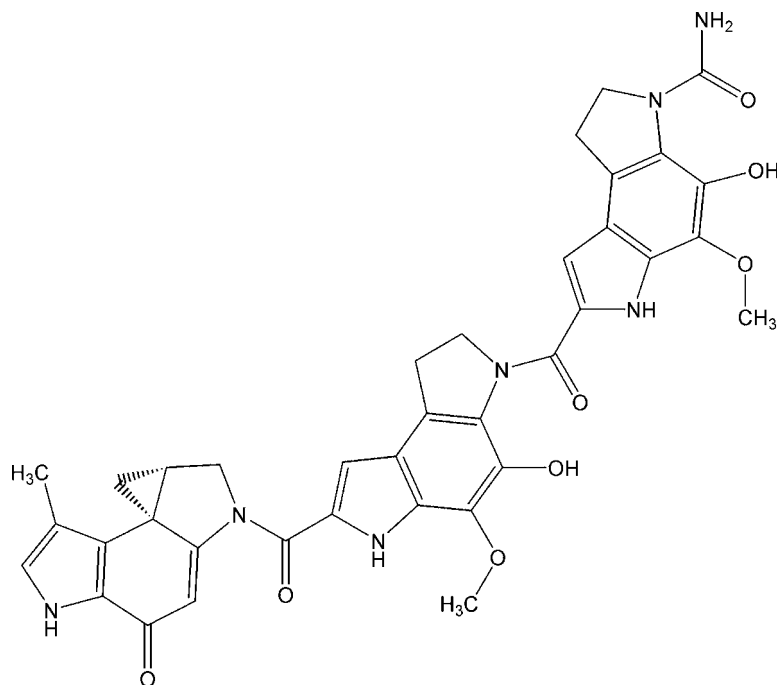


Figure 5.61 The structure of the covalently binding drug CC-1065.

was rationally designed to covalently cross-link to the hexamer sequence 5'-TAATTA. This sequence is intrinsically bent in solution, as a consequence of the high propeller twist of the A•T base pairs. Unexpectedly, NMR studies (Thompson and Hurley, 1995) have shown that there are two adduct conformations, in a 40:60 ratio. Both have straightened-out A-tracts, but differ in the hydrogen-bonding arrangement of the central A•T base pairs, with one being unpaired and the other having a Hoogsteen arrangement.

Minor-groove alkylation most often occurs at the N2 of guanine. A number of compounds that bind at this site have significant experimental antitumour activity. Mitomycin, discovered in Japan in fermentation broths over 50 years ago, has found some clinical application in the treatment of solid tumors, although its severe toxicity has precluded extensive use. A few other minor-groove alkylators have now progressed to evaluation in clinical trials. Their biological efficacy is undoubtedly a consequence of interference with particular aspects of gene regulation, although this is as yet poorly understood. For at least some compounds such as DSB-120 and its analogues (see below), their binding to the minor groove of B-form DNA involves little if any perturbation from the canonical structure. This results in a lack of structural lesions for subsequent recognition by a repair protein, so that these adducts tend to be long-lasting, with consequent enhancement of potency.

The pyrrolo[2,1-c][1,4]benzodiazepine (PBD) family of compounds, typified by the naturally occurring antitumor antibiotic anthramycin, bind to the exocyclic N2 atom of guanine bases through the N10-C11 double bond (Fig. 5.62). The crystal structure (Kopka *et al.*, 1994) of an adduct with two bound drug molecules (Fig. 5.63) shows that each lies in the minor groove, in a manner analogous to that of the

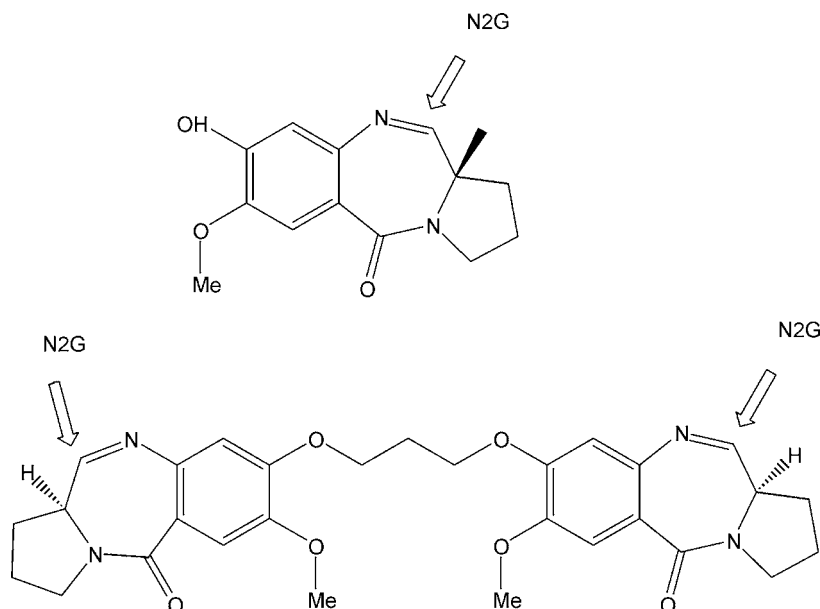


Figure 5.62 The drug anthramycin (top), showing its point of attachment to the exocyclic amino group of guanine. The anthramycin dimer DSB-120 is shown below.

noncovalent binders such as berenil and netropsin. The binding site for anthramycin is three bases long, with a preference for 5'-Pu-G-Pu type sequences. Several dimers of these molecules have been designed which have exceptional cytotoxicity. The best-studied of these is DSB-120 (Bose *et al.*, 1992), with an IC_{50} of $0.0005\ \mu\text{M}$, making it some 1000 times more potent than the equivalent monomer in some cell lines. These cell kill effects are probably due to inhibition of transcription, although it is not clear to what extent sequence-selective effects are involved. DSB-120 (Fig. 5.62) has a site size requirement of 6–7 base pairs, with a sequence preference for sites of the type 5'-Pu-GATC-Py. Its ability to form interstrand covalent cross-links to two sites in the minor groove, together with its almost ideal isohelicity, result in a tight fit to canonical B-form DNA. NMR and molecular modelling studies have shown that DSB-120 produces almost no structural perturbation to DNA upon cross-linking. This feature of the PDB dimer complexes is undoubtedly significant for their resistance to DNA repair and may contribute to the high *in vivo* antitumor activity shown by several members of this series, one of which, SJG-131, has now proceeded to clinical trial evaluation (Alley *et al.*, 2004).

Mitomycin C requires metabolic activation in order to react with DNA (Tomasz and Palom, 1997). Opening of the aziridine ring (Fig. 5.64) results in covalent attachment to one guanine in the sequence 5'-CpG. A second is through the carbamate group. The structures of both this cross-linked adduct and a mono-adduct from aziridine ring-opening, have been characterized by NMR methods (Norman *et al.*, 1990; Sastry *et al.*, 1995), as has the very different N7 adduct, which places the bound drug in the DNA major groove (Subramanian *et al.*, 2001), without significant perturbations from B-DNA form.



Figure 5.63 The crystal structure (Kopka *et al.*, 1994) of an oligonucleotide duplex complex with two covalently bound anthramycin molecules, highlighted in black.

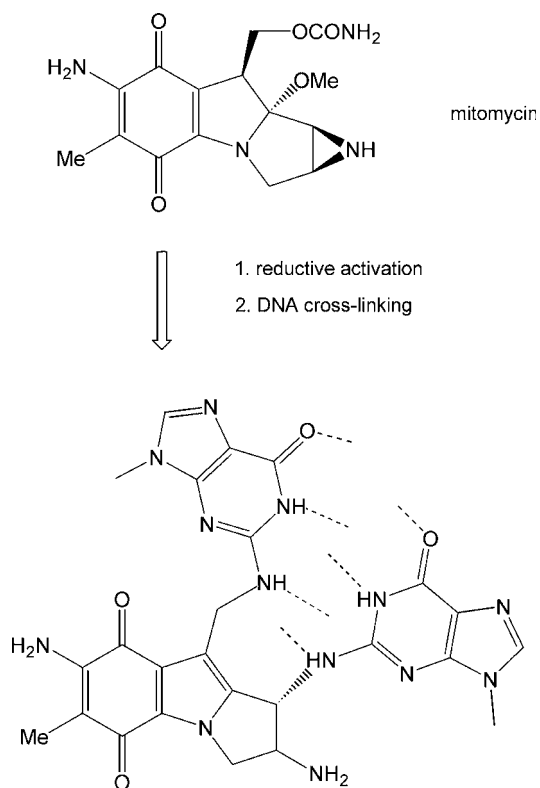


Figure 5.64 The structure of mitomycin, together with a schematic diagram showing a DNA-mitomycin adduct cross-linked to two guanine bases on the opposite strands of a DNA duplex.

High DNA sequence selectivity is a dominant feature of several of these minor-groove covalent-binding molecules. This, together with their druglike characteristics, does suggest that a fruitful future direction for the development of therapeutic sequence-selectivity at the gene level, may be in combining such molecules with the features of noncovalent recognition such as have been described earlier in this chapter.

References

- Abu-Daya, A., Brown, P. M., and Fox, K. R. (1995). DNA sequence preferences of several AT-selective minor groove binding ligands. *Nucleic Acids Res.*, **17**, 3385–3392.
- Adams, A., Guss, J. M., Collyer, C. A., Denny, W. A., and Wakelin, L. P. G. (2000). A novel form of intercalation involving four DNA duplexes in an acridine-4-carboxamide complex of d(CGTACG)₂. *Nucleic Acids Res.*, **28**, 4244–4253.
- Adams, A., Guss, J. M., Collyer, C. A., Denny, W. A., Prakash, A. S., and Wakelin, L. P. G. (2000). Acridinecarboxamide topoisomerase poisons: structural and kinetic studies of the DNA complexes of 5-substituted 9-amino-(N-(2-dimethylamino)ethyl)acridine-4-carboxamides. *Mol. Pharmacol.*, **58**, 649–658.
- Alden, C. J. and Kim, S.-H. (1979). Solvent-accessible surfaces of nucleic acids. *J. Mol. Biol.*, **132**, 411–434.
- Alley, M. C., Hollingshead, M. G., Pacula-Cox, C. M., Waud, W. R., Hartley, J. A., Howard, P. W., Gregson, S. J., Thurston, D. E., and Sausville, E. A. (2004). SJG-136 (NSC 694501): a novel rationally designed DNA minor groove interstrand cross-linking agent with potent and broad spectrum antitumor activity: part 2: efficacy evaluations. *Cancer Res.*, **64**, 6700–6706.
- Bailly, C., Chessare, G., Carrasco, C., Joubert, A., Mann, J., Wilson, W. D., and Neidle, S. (2003). Sequence-specific minor groove binding by bis-benzimidazoles: water molecules in ligand recognition. *Nucleic Acids Res.*, **31**, 1514–1524.
- Bennett, M., Krah, A., Wien, F., Garman, E., McKenna, R., Sanderson, M., and Neidle, S. (2000). A DNA-porphyrin minor-groove complex at atomic resolution: the structural consequences of porphyrin ruffling. *Proc. Natl. Acad. Sci. USA*, **97**, 9476–9481.
- Billeter, M., Güntert, P., Luginbühl, P., and Wüthrich, K. (1996). Hydration and DNA recognition by homeodomains. *Cell*, **85**, 1057–1065.
- Bose, D. S., Thompson, A. S., Ching, J., Hartley, J. A., Berardini, M. D., Jenkins, T. C., Neidle, S., Hurley, L. H., and Thurston, D. E. (1992). Rational design of a highly efficient irreversible DNA inter-strand cross-linking agent based on the pyrrolobenzodiazepine ring system. *J. Amer. Chem. Soc.*, **114**, 4939–4941.
- Bostock-Smith, C. E., Laughton, C. A., and Searle, M. S. (1999). Solution structure and dynamics of the A-T tract DNA decamer duplex d(GGTAATTACC)₂: implications for recognition by minor groove binding drugs. *Biochem. J.*, **342**, 125–132.
- Boykin, D. W., Kumar, A., Spychala, J., Zhou, M., Lombardy, R. J., Wilson, W. D., Dykstra, C. C., Jones, S. K., Hall, J. E., Tidwell, R. R., Laughton, C., Nunn, C. M., and Neidle, S. (1995). Dicationic diarylfurans as anti-Pneumocystis carinii agents. *J. Med. Chem.*, **38**, 912–916.
- Bremer, R. E., Wurtz, N. R., Szewczyk, J. W., and Dervan, P. B. (2001). Inhibition of major groove DNA binding bZIP proteins by positive patch polyamides. *Bioorg. Med. Chem.*, **9**, 2093–2103.
- Brogden, A. L., Hopcroft, N. H., Searcey, M., and Cardin, C. J. (2007). Ligand bridging of the DNA Holliday junction: molecular recognition of a stacked X four-way junction by a small molecule. *Angew. Chemie. Int. Edit.*, **46**, 3850–3854.
- Brown, D. G., Sanderson, M. R., Garman, E., and Neidle, S. (1992). Crystal structure of a berenil-d(CGCAAATTTGCG) complex. An example of drug-DNA recognition based on sequence-dependent structural features. *J. Mol. Biol.*, **226**, 481–490.
- Burnett, R., Melnder, C., Puckett, J. W., Son, L. S., Wells, R. D., Dervan, P. B., and Gottesfeld, J. M. (2006). DNA sequence-specific polyamides alleviate transcription inhibition associated with long GAA•TTC repeats in Friedreich's ataxia. *Proc. Natl. Acad. Sci. USA*, **103**, 11497–11502.
- Canals, A., Purciolas, M., Aymami, J., and Coll, M. (2005). The anticancer agent ellipticine unwinds DNA by intercalative binding in an orientation parallel to base pairs. *Acta Crystallogr.*, **D61**, 1009–1012.
- Cassidy, S. A., Strekowski, L., Wilson, W. D., and Fox, K. R. (1994). Effect of a triplex-binding ligand on parallel and antiparallel DNA triple helices using short unmodified and acridine-linked oligonucleotides. *Biochemistry*, **33**, 15338–15347.
- Cech, T. R. (2000). Life at the end of the chromosome: telomeres and telomerase. *Ang. Chem. Internat. Edit.*, **39**, 34–43.

- Chaires, J. B., Leng, F., Przewloka, T., Fokt, I., Ling, Y.-H., Perez-Soler, R., and Priebe, W. (1997). Structure-based design of a new bisintercalating anthracycline antibiotic. *J. Med. Chem.*, **40**, 261–266.
- Chen, X., Ramakrishnan, B., Rao, S. T., and Sundaralingam, M. (1994). Binding of two distamycin A molecules in the minor groove of an alternating B-DNA duplex. *Nature Struct. Biol.*, **1**, 169–175.
- Chiu, T. K., Kaczor-Grzeskowiak, M., and Dickerson, R. E. (1999). Absence of minor groove monovalent cations in the crosslinked dodecamer C-G-C-G-A-A-T-T-C-G-C-G. *J. Mol. Biol.*, **292**, 589–608.
- Cieplak, P., Rao, S. N., Grootenhuis, P. D., and Kollman, P. A. (1990). Free energy calculation on base specificity of drug–DNA interactions: application to daunomycin and acridine intercalation into DNA. *Biopolymers*, **29**, 717–727.
- Clark, G. R., Gray, E. J., Neidle, S., Ji, Y.-H., and Leupin, W. (1996). Isohelicity and phasing in drug–DNA sequence recognition: crystal structure of a tris(benzimidazole)-oligonucleotide complex. *Biochemistry*, **35**, 13745–13752.
- Coll, M., Frederick, C. A., Wang, A. H.-J., and Rich, A. (1987). A bifurcated hydrogen-bonded conformation in the d(A.T) base pairs of the DNA dodecamer d(CGCAAATTTGCG) and its complex with distamycin. *Proc. Natl. Acad. Sci. USA*, **84**, 8385–8389.
- Coste, F., Malinge, J.-M., Serre, L., Shepard, W., Roth, M., Leng, M., and Zelwer, C. (1999). Crystal structure of a double-stranded DNA containing a cisplatin interstrand cross-link at 1.63 Å resolution: hydration at the platinum site. *Nucleic Acids Res.*, **27**, 1837–1846.
- Cuesta-Seijo, J. A. and Sheldrick, G. M. (2005). Structures of complexes between echinomycin and duplex DNA. *Acta Crystallogr.*, **D61**, 442–448.
- Cuesta-Seijo, J. A., Weiss, M. S. and Sheldrick, G. M. (2006). Serendipitous SAD phasing of an echinomycin-(ACGTACGT)₂ bisintercalation complex. *Acta Crystallogr.*, **D62**, 417–424.
- de Clairac, R. P. L., Geierstanger, B. H., Mrksich, M., Dervan, P. B., and Wemmer, D. E. (1997). NMR characterization of hairpin polyamide complexes with the minor groove of DNA. *J. Amer. Chem. Soc.*, **119**, 7909–7916.
- Dervan, P. B. (1986). Design of sequence-specific DNA-binding molecules. *Science*, **232**, 464–471.
- Dervan, P. B. and Bürl, R. W. (1999). Sequence-specific DNA recognition by polyamides. *Curr. Opin. Struct. Biol.*, **3**, 688–693.
- Dickinson, L. A., Trauger, J. W., Eldon, E. B., Ghazal, P., Dervan, P. B., and Gottesfeld, J. M. (1999). Anti-repression of RNA polymerase II transcription by pyrrole-imidazole polyamides. *Biochemistry*, **38**, 10801–10807.
- Dolenc, J., Oostenbrink, C., Koller, J., and van Gunsteren, W. F. (2005). Molecular dynamics simulations and free energy calculations of netropsin and distamycin binding to an AAAAA DNA binding site. *Nucleic Acids Res.*, **33**, 725–733.
- Drew, H. R. and Dickerson, R. E. (1981). Structure of a B-DNA dodecamer: III. Geometry of hydration. *J. Mol. Biol.*, **151**, 535–556.
- Duan, Y., Wilkosz, P., Crowley, M., and Rosenberg, J. M. (1997). Molecular dynamics simulation study of DNA dodecamer d(CGCGAATTCGCG) in solution: conformation and hydration. *J. Mol. Biol.*, **272**, 553–572.
- Egli, M., Tereshko, V., Teplova, M., Minasov, G., Joachimiak, A., Sanishvili, R., Weeks, C. M., Miller, R., Maier, M. A., Ari, H., Cook, P. D., and Manoharan, M. (1998). X-ray crystallographic analysis of the hydration of A- and B-form DNA at atomic resolution. *Biopolymers* **48**, 234–252.
- Egli, M., Williams, L. D., Frederick, C. A., and Rich, A. (1991). DNA-nogalamycin interactions. *Biochemistry*, **30**, 1364–1372.
- Eichman, B. F., Mooers, B. H. M., Alberti, M., Hearst, J. E., and Ho, P. S. (2001). The crystal structures of psoralen cross-linked DNAs: drug-dependent formation of Holliday junctions. *J. Mol. Biol.*, **308**, 15–26.
- Eis, P. S., Smith, J.-A., Rydzewski, J. M., Chase, D. A., Boger, D. L., and Chazin, W. J. (1997). High resolution solution structure of a DNA duplex alkylated by the antitumor agent duocarmycin SA. *J. Mol. Biol.*, **272**, 237–252.
- Escude, C., Nguyen, C. H., Kukreti, S., Janin, Y., Sun, J. S., Bisagni, E., Garestier, T., and Hélène, C. (1998). Rational design of a triple helix-specific intercalating ligand. *Proc. Natl. Acad. Sci. USA*, **95**, 3591–3596.
- Evans, D. A. and Neidle, S. (2006). Virtual screening of DNA minor groove binders. *J. Med. Chem.*, **49**, 4232–4238.
- Fedoroff, O. Y., Salazar, M., Han, H., Chemeris, V. V., Kerwin, S. M., and Hurley, L. H. (1998). NMR-based model of a telomerase-inhibiting compound bound to G-quadruplex DNA. *Biochemistry*, **37**, 12367–12374.
- Feig, M. and Pettitt, B. M. (1999). Modeling high-resolution hydration patterns in correlation with DNA sequence and conformation. *J. Mol. Biol.*, **286**, 1075–1095.
- Fox, K. R. and Waring, M. J. (1986). Nucleotide sequence binding preferences of nogalamycin investigated by DNase I footprinting. *Biochemistry*, **25**, 4349–4356.
- Fox, K. R., Polucci, P., Jenkins, T. C., and Neidle, S. (1995). A molecular anchor for stabilizing triple-helical DNA. *Proc. Natl. Acad. Sci. USA*, **92**, 7887–7891.

- Fuller, W., Forsyth, T., and Mahendrasingam, A. (2004). Water-DNA interactions as studied by X-ray and neutron fibre diffraction. *Phil. Trans. Roy. Soc. Lond. B.*, **359**, 1237–1247.
- Gao, X., Mirau, P., and Patel, D. J. (1992). Structure refinement of the chromomycin dimer-DNA oligomer complex in solution. *J. Mol. Biol.*, **223**, 259–279.
- Gao, X. and Patel, D. J. (1988). NMR studies of echinomycin bisintercalation complexes with d(A1-C2-G3-T4) and d(T1-C2-G3-A4) duplexes in aqueous solution: sequence-dependent formation of Hoogsteen A1.T4 and Watson–Crick T1.A4 base pairs flanking the bisintercalation site. *Biochemistry*, **27**, 1744–1751.
- Gavathiotis, E., Heald, R. A., Stevens, M. F., and Searle, M. S. (2003). Drug recognition and stabilisation of the parallel-stranded DNA quadruplex d(TTAGGGT)₄ containing the human telomeric repeat. *J. Mol. Biol.*, **334**, 25–36.
- Gavathiotis, E., Sharman, G. J., and Searle, M. S. (2000). Sequence-dependent variation in DNA minor groove width dictates orientational preference of Hoechst 33258 in A-tract recognition: solution NMR structure of the 2:1 complex with d(CTTTTGCAAAAG)₂. *Nucleic Acids Res.*, **28**, 728–735.
- Ge, W., Schneider, B., and Olson W. K. (2005). Knowledge-based elastic potentials for docking drugs or proteins with nucleic acids. *Biophysical J.*, **88**, 1166–1190.
- Gearhart, M. D., Dickinson, L., Ehley, J., nelander, C., Dervan, P. B., Wright, P. E., and Gottesfeld, J. M. (2005). Inhibition of DNA binding by human estrogen-related receptor 2 and estrogen receptor alpha with minor groove binding polyamides. *Biochemistry*, **44**, 4196–4203.
- Gelasco, A. and Lippard, S. J. (1998). NMR Solution structure of a DNA dodecamer duplex containing a *cis*-diammineplatinum(II) d(GpG) intrastrand cross-link, the major adduct of the anticancer drug cisplatin. *Biochemistry*, **37**, 9230–9239.
- Gilbert, D. E. and Feigon, J. (1992). Proton NMR study of the [d(ACGTATACGT)]₂-echinomycin complex: conformational changes between echinomycin binding sites. *Nucleic Acids Res.*, **20**, 2411–2420.
- Gochin, M. (2000). A high-resolution structure of a DNA-chromomycin-Co(II) complex determined from pseudo-contact shifts in nuclear magnetic resonance. *Structure*, **8**, 441–452.
- Goodsell, D. and Dickerson, R. E. (1986). Isohelical analysis of DNA groove-binding drugs. *J. Med. Chem.*, **29**, 727–733.
- Goodwin, K. D., Lewis, M. A., Tanius, F. A., Tidwell, R. R., Wilson, W. D., Georgiadis, M. M., and Long, E. C. (2006). A high-throughput, high-resolution strategy for the study of site-selective DNA binding agents: Analysis of a “highly twisted” benzimidazole-diamidine. *J. Amer. Chem. Soc.*, **126**, 7846–7854.
- Goodwin, K. D., Long, E. C., and Georgiadis, M. M. (2005). A host-guest approach for determining drug-DNA interactions: an example using netropsin. *Nucleic Acids Res.*, **33**, 4106–4116.
- Gottesfeld, J. M., Melander, C., Suto, R. K., Raviol, H., Luger, K., and Dervan, P. B. (2001). Sequence-specific recognition of DNA in the nucleosome by pyrrole-imidazole polyamides. *J. Mol. Biol.*, **309**, 615–629.
- Gottesfeld, J. M., Neely, L., Trauger, J. W., Baird, E. E., and Dervan, P. B. (1997). Regulation of gene expression by small molecules. *Nature*, **387**, 202–205.
- Guo, Q., Lu, M., Marky, L. A., and Kallenbach, N. R. (1992). Interaction of the dye ethidium bromide with DNA containing guanine repeats. *Biochemistry*, **31**, 2451–2455.
- Haider, S. M., Parkinson, G. N., and Neidle, S. (2003). Structure of a G-quadruplex-ligand complex. *J. Mol. Biol.*, **326**, 117–125.
- Halle, B. and Denisov, V. P. (1998). Water and monovalent ions in the minor groove of B-DNA oligonucleotides as seen by NMR. *Biopolymers*, **48**, 210–233.
- Han, H. and Hurley, L. H. (2000). G-quadruplex DNA: a potential target for anti-cancer drug design. *Trends Pharm. Sci.*, **21**, 136–142.
- Hartley, J. A., Bingham, J. P., and Souhami, R. L. (1992). DNA sequence selectivity of guanine-N7 alkylation by nitrogen mustards is preserved in intact cells. *Nucleic Acids Res.*, **20**, 3175–3178.
- Henikoff, S. and Vermaak, D. (2000). Bugs on drugs go GAGAA. *Cell*, **103**, 695–698.
- Hopcroft, N. H., Brogden, A. L., Searcey, M., and Cardin, C. J. (2006). X-ray crystallographic study of DNA duplex cross-linking: simultaneous binding to two d(CGTACG)₂ molecules by a bis(9-aminoacridine-4-carboxamide) derivative. *Nucleic Acids Res.*, **34**, 6663–6672.
- Hou, M.-H., Robinson, H., Gao, Y.-G., and Wang, A. H.-J. (2004). Crystal structure of the [Mg²⁺-(chromomycin A₃)₂]-d(TTGGCCAA)₂ complex reveals GGCC binding specificity of the drug dimer chelated by a metal ion. *Nucleic Acids Res.*, **32**, 2214–2222.
- Hu, G. G., Shui, X., Leng, F., Priebe, W., Chaires, J. B., and Williams, L. D. (1997). Structure of a DNA-bisdaunomycin complex. *Biochemistry*, **36**, 5940–5946.
- Huang, H., Zhu, L., Reid, B. R., Drobny, G. P., and Hopkins, P. B. (1995). Solution structure of a cisplatin-induced DNA interstrand cross-link. *Science*, **270**, 1842–1845.
- Hud, N. V., Sklenár, V., and Feigon, J. (1999). Localization of ammonium ions in the minor groove of DNA duplexes in solution and the origin of DNA A-tract bending. *J. Mol. Biol.*, **286**, 651–660.

- Hurley, L. H., Lee, C.-S., McGovren, J. P., Mitchell, M. A., Warpehoski, M. A., Kelly, R. C., and Aristoff, P. A. (1988). Molecular basis for sequence-specific DNA alkylation by CC-1065. *Biochemistry*, **27**, 3886–3892.
- Janssen, S., Cuvier, O., Müller, M., and Laemmli, U. K. (2000). Specific gain- and loss-of-function phenotypes induced by satellite-specific DNA-binding drugs fed to *Drosophila melanogaster*. *Mol. Cell*, **6**, 1013–1024.
- Janssen, S., Durussel, T., and Laemmli, U. K. (2000). Chromatin opening of DNA satellites by targeted sequence-specific drugs. *Mol. Cell*, **6**, 999–1011.
- Ji, Y.-H., Bur, D., Häslér, W., Schmitt, V. R., Dorn, A., Bailly, C., Waring, M. J., Hochstrasser, R., and Leupin, W. (2001). Tris-benzimidazole derivatives: design, synthesis and DNA sequence recognition. *Bioorg. Med. Chem.*, **9**, 2905–2919.
- Johansson, E., Parkinson, G., and Neidle, S. (2000). A new crystal form for the dodecamer C-G-C-G-A-A-T-T-C-G-C-G: symmetry effects on sequence-dependent DNA structure. *J. Mol. Biol.*, **300**, 551–561.
- Kamitori, S. and Takusagawa, F. (1992). Crystal structure of the 2:1 complex between d(GAAGCTTC) and the anti-cancer drug actinomycin D. *J. Mol. Biol.*, **225**, 445–456.
- Kang, S., Shafer, R. H., and Kuntz, I. D. (2004). Calculation of ligand-nucleic acid binding free energies with the generalized-born model in DOCK. *Biopolymers*, **73**, 192–204.
- Kennard, O., Cruse, W. B. T., Nachman, J., Prange, T., Shakked, Z., and Rabinovich, D. (1986). Ordered water structure in an A-DNA octamer at 1.7 Å resolution. *J. Biomol. Struct. Dynamics*, **3**, 623–647.
- Keppler, M. D., Read, M. A., Perry, P. J., Trent, J. O., Jenkins, T. C., Reszka, A. P., Neidle, S., and Fox, K. R. (1999). Stabilization of DNA triple helices by a series of mono- and disubstituted amidoanthraquinones. *Eur. J. Biochem.*, **263**, 817–825.
- Kielkopf, C. L., Baird, E. E., Dervan, P. B., and Rees, D. C. (1998a). Structural basis for G.C recognition in the DNA minor groove. *Nature Struct. Biol.*, **5**, 104–109.
- Kielkopf, C. L., Bremer, R. E., White, S., Szewczyk, J. W., Turner, J. M., Baird, E. E., Dervan, P. B., and Rees, D. C. (2000). Structural effects of DNA sequence on T.A recognition by hydroxypyrrole/pyrrole pairs in the minor groove. *J. Mol. Biol.*, **295**, 557–567.
- Kielkopf, C. L., Erkkila, K. E., Hudson, B. P., Barton, J. K., and Rees, D. C. (2000). Structure of a photoactive rhodium complex intercalated into DNA. *Nature Struct. Biol.*, **7**, 117–121.
- Kielkopf, C. L., White, S., Szewczyk, J. W., Turner, J. M., Baird, E. E., Dervan, P. B., and Rees, D. C. (1998b). A structural basis for recognition of A.T and T.A base pairs in the minor groove of B-DNA. *Science*, **282**, 111–115.
- Kim, M.-Y., Vankayalapati, H., Shin-ya, K., Wierzb, K., and Hurley, L. H. (2002). Telomestatin, a potent telomerase inhibitor that interacts quite specifically with the human telomeric intramolecular G-Quadruplex. *J. Amer. Chem. Soc.*, **124**, 2098–2099.
- Koeppel, F., Riou, J.-F., Laoui, A., Mailliet, P., Arimondo, P. B., Labit, D., Petitgenet, O., Hélène, C., and Mergny, J.-L. (2001). Ethidium derivatives bind to G-quartets, inhibit telomerase and act as fluorescent probes for quadruplexes. *Nucleic Acids Res.*, **29**, 1087–1096.
- Kopka, M. L., Goodsell, D. S., Baikalov, I., Grzeskowiak, K., Cascio, D., and Dickerson, R. E. (1994). Crystal structure of a covalent DNA-drug adduct: anthramycin bound to C-C-A-A-C-G-T-T-G-G and a molecular explanation of specificity. *Biochemistry*, **33**, 13593–13610.
- Kopka, M. L., Goodsell, D. S., Han, G. W., Chiu, T. K., Lown, J. W., and Dickerson, R. E. (1997). Defining GC-specificity in the minor groove: side-by-side binding of the di-imidazole lexitropsin to C-A-T-G-G-C-C-A-T-G. *Structure*, **5**, 1033–1046.
- Kopka, M. L., Yoon, C., Goodsell, D., Pjura, P., and Dickerson, R. E. (1985). Binding of an antitumor drug to DNA, Netropsin and C-G-C-G-A-A-T-T-BrC-G-C-G. *J. Mol. Biol.*, **183**, 553–563.
- Kubinec, M. G. and Wemmer, D. E. (1992). NMR evidence for DNA bound water in solution. *J. Amer. Chem. Soc.*, **114**, 8739–8740.
- Lane, A. N., Jenkins, T. C., Brown, T., and Neidle, S. (1991). Interaction of berenil with the EcoRI dodecamer d(CGCGAATTCGCG)₂ in solution studied by NMR. *Biochemistry*, **30**, 1372–1385.
- Laughton, C. and Luisi, B. (1998). The mechanics of minor groove width variation in DNA, and its implications for the accommodation of ligands. *J. Mol. Biol.*, **288**, 953–963.
- Laughton, C. A., Tanious, F. A., Nunn, C. M., Boykin, D. W., Wilson, W. D., and Neidle, S. (1996). A crystallographic and spectroscopic study of the complex between d(CGCGAATTCGCG)₂ and 2,5-bis(4-guanyphenyl)furan, an analogue of berenil. Structural origins of enhanced DNA-binding affinity. *Biochemistry*, **35**, 5655–5661.
- Lerman, L. S. (1961). Structural considerations in the interaction of DNA and acridines. *J. Mol. Biol.*, **3**, 18–30.
- Liepinsh, E., Otting, G., and Wüthrich, K. (1992). NMR observation of individual molecules of hydration water bound to DNA duplexes: direct evidence for a spine of hydration water present in aqueous solution. *Nucleic Acids Res.*, **20**, 6549–6553.
- Lipscomb, L. A., Zhou, F. X., Presnell, S. R., Woo, R. J., Peek, M. E., Plaskon, R. R., and Williams, L. D. (1996). Structure of a DNA-porphyrin complex. *Biochemistry*, **35**, 2818–2823.

- Liu, J. and Subirana, J. (1999). Structure of d(CGCGAATTCGCG) in the presence of Ca^{2+} ions. *J. Biol. Chem.*, **274**, 24749–24752.
- Liu, X., Chen, H., and Patel, D. J. (1991). Solution structure of actinomycin-DNA complexes: drug intercalation at isolated G-C sites. *J. Biomol. NMR*, **1**, 323–347.
- Lown, J. W. (1994). DNA recognition by lexitropsins, minor groove binding agents. *J. Mol. Recognition*, **7**, 79–88.
- Maeshima, K., Janssen, S., and Laemmli, U. K. (2001). Specific targeting of insect and vertebrate telomeres with pyrrole and imidazole polyamides. *EMBO J.*, **20**, 3218–3228.
- Mapp, A. K., Ansari, A. Z., Ptashne, M., and Dervan, P. B. (2000). Activation of gene expression by small molecule transcription factors. *Proc. Natl. Acad. Sci. USA*, **97**, 3930–3935.
- Marques, M. A., Doss, R. M., Foister, S., and Dervan P. B. (2004). Expanding the repertoire of heterocycle ring pairs for programmable minor groove DNA recognition. *J. Amer. Chem. Soc.*, **126**, 10339–10349.
- Marzilli, L. G., Banville, and D. L., Zon, G. (1986). Pronounced proton and phosphorus-31 NMR spectral changes on meso-tetrakis(N-methylpyridinium-4-yl)porphyrin binding to poly[d(G-C)].poly[d(G-C)] and to three tetra-decaoligodeoxyribonucleotides: evidence for symmetric, selective binding to 5'CG3' sequences. *J. Amer. Chem. Soc.*, **108**, 4188–4192.
- McConnell, K. J. and Beveridge, D. L. (2000). DNA Structure: What's in charge? *J. Mol. Biol.*, **304**, 803–820.
- McEachern, M. J., Krauskopf, A., and Blackburn, E. H. (2000). Telomeres and their control. *Ann. Rev. Genetics*, **34**, 331–358.
- McLean, M. J., Seela, F., and Waring, M. J. (1989). Echinomycin-induced hypersensitivity to osmium tetroxide of DNA fragments incapable of forming Hoogsteen base pairs. *Proc. Natl. Acad. Sci. USA*, **86**, 9687–9691.
- Mergny, J.-L., Lacroix, L., Teulado-Fichou, M.-P., Hounsou, C., Guittat, L., Hoarau, M., Arimondo, P. B., Vigneron, J.-P., Lehn, J.-M., Riou, J.-F., Garestier, T., and Hélène, C. (2001). Telomerase inhibitors based on quadruplex ligands selected by a fluorescence assay. *Proc. Natl. Acad. Sci. USA*, **98**, 3062–3067.
- Mergny, J.-L. and Hélène, C. (1998). G-quadruplex DNA: a target for drug design. *Nature Medicine*, **4**, 1366–1367.
- Mergny, J.-L., Duval-Valentin, G., Nguyen, C. H., Perrouault, L., Faucon, B., Rougée, Montenay-Garestier, T., Bisagni, E., and Hélène, C. (1992). Triple helix-specific ligands. *Science*, **256**, 1681–1684.
- Miao, Y. Lee, M. P. H., Parkinson, G. N., Batista-Parra, A., Ismail, M., Neidle, S., Boykin, D. W., and Wilson W. D. (2005). Out-of-shape DNA minor groove binders: induced fit interactions of heterocyclic dictations with the DNA minor groove. *Biochemistry* **44**, 14701–14708.
- Minasov, G., Tereshko, V., and Egli, M. (1999). Atomic-resolution crystal structures of B-DNA reveal specific influences of divalent metal ions on conformation and packing. *J. Mol. Biol.*, **291**, 83–99.
- Mrksich, M. and Dervan, P. B. (1993). Antiparallel side-by-side heterodimer for sequence-specific recognition in the minor groove of DNA by a distamycin/1-methylimidazole-2-carboxamide-netropsin pair. *J. Amer. Chem. Soc.*, **115**, 2572–2576.
- Neidle, S., Achari, A., Taylor, G. L., Berman, H. M., Carrell, H. L., Glusker, J. P., and Stallings, W. C. (1977). Structure of a dinucleoside phosphate-drug complex as model for nucleic acid–drug interaction. *Nature*, **269**, 304–307.
- Norman, D., Live, D., Sastry, M., Lipman, R., Hingerty, B. E., Tomasz, M., Broyde, S., and Patel, D. J. (1990). NMR and computational characterization of mitomycin cross-linked to adjacent deoxyguanosines in the minor groove of the d(T-A-C-G-T-A).d(T-A-C-G-T-A) duplex. *Biochemistry* **29**, 2861–2875.
- Nguyen, B., Hamelberg, D., Bailly, C., Colson, P., Stanek, J., Brun, R., Neidle, S., and Wilson, W. D. (2004). Characterization of a novel DNA minor-groove complex. *Biophysical J.*, **86**, 1028–1041.
- Nguyen, B., Lee, M. P. H., Hamelberg, D., Joubert, A., Bailly, C., Brun, R., Neidle, S., and Wilson, W. D. (2002). Strong binding in the DNA minor groove by an aromatic diamidine with a shape that does not match the curvature of the groove. *J. Amer. Chem. Soc.*, **124**, 13680–13681.
- Nunn, C. M., Garman, E., and Neidle, S. (1997). Crystal structure of the DNA decamer d(CGCAATTGCG) complexed with the minor groove binding drug netropsin. *Biochemistry*, **36**, 4792–4799.
- Ohndorf, U.-M., Rould, M. A., He, Q., Pabo, C. O., and Lippard, S. J. (1999). Basis for recognition of cis-platinum-modified DNA by high-mobility-group proteins. *Nature*, **399**, 708–712.
- Oleksi, A., Blanco, A. G., Boer, R., Usón, I., Aymani, J., Rodger, A., Hannon, J. J., and Coll M. (2006). Molecular recognition of a three-way DNA junction by a metallosupramolecular helicate. *Angew. Chem. Int. Ed.*, **45**, 1227–1231.
- Orgenesian, L. and Bryan, T. M. (2007). Physiological relevance of telomeric G-quadruplex formation: a potential drug target. *Bioessays*, **29**, 155–165.
- Pal, S. K., Zhao, L., Xia, T., and Zewail, A. H. (2003). Site- and sequence-selective ultrafast hydration of DNA. *Proc. Natl. Acad. Sci. USA*, **100**, 13746–13751.
- Parkinson, G. N., Ghosh, R., and Neidle, S. (2007). Structural basis for binding of porphyrin to human telomeres. *Biochemistry*, **46**, 2390–2397.

- Patel, D. J. and Shapiro, L. (1986). Sequence-dependent recognition of DNA duplexes. Netropsin complexation to the AATT site of the d(G-G-A-A-T-T-C-C) duplex in aqueous solution. *J. Biol. Chem.*, **261**, 1230–1240.
- Peek, M. E., Lipscomb, L. A., Bertrand, J. A., Gao, Q., Roques, B. P., Garbay-Jaureguiberry, C., and Williams, L. D. (1994). DNA distortion in bis-intercalated complexes. *Biochemistry*, **33**, 3794–3800.
- Pelton, J. G. and Wemmer, D. E. (1990). Binding modes of distamycin A with d(CGCAAATTTGCG)₂ determined by two-dimensional NMR. *J. Amer. Chem. Soc.*, **112**, 1393–1399.
- Phan, A. T., Kuryavii, V., Gaw, H. Y., and Patel, D. J. (2005). Small-molecule interaction with a five-guanine-tract G-quadruplex structure from the human MYC promoter. *Nature Chem. Biol.*, **1**, 167–173.
- Phan, A. T., Leroy, J.-L., and Guéron, M. (1999). Determination of the residence time of water molecules hydrating B'-DNA and B-DNA, by one-dimensional zero-enhancement nuclear Overhauser effect spectroscopy. *J. Mol. Biol.*, **286**, 505–519.
- Pierre, V. C., Kaiser, J. T., and Barton J. K. (2007). Insights into finding a mismatch through the structure of a mispaired DNA bound by a rhodium intercalator. *Proc. Natl. Acad. Sci. USA*, **104**, 429–434.
- Pullman, A. and Pullman, B. (1981). Molecular electrostatic potential of the nucleic acids. *Quart. Rev. Biophys.*, **14**, 289–380.
- Quigley, G. J., Ughetto, G., van der Marel, G. A., van Boom, J. H., Wang, A. H.-J., and Rich, A. (1986). Non-Watson-Crick G.C and A.T base pairs in a DNA-antibiotic complex. *Science*, **232**, 1255–1258.
- Rangan, A., Fedoroff, O. Y., and Hurley, L. H. (2001). Induction of duplex to G-quadruplex transition in the c-myc promoter region by a small molecule. *J. Biol. Chem.*, **276**, 4640–4646.
- Read, M. A., Wood, A. A., Harrison, J. R., Gowan, S. M., Kelland, L. R., Dosanjh, H. S., and Neidle, S. (1999). Molecular modeling studies on G-quadruplex complexes of telomerase inhibitors: structure-activity relationships. *J. Med. Chem.*, **42**, 4538–4546.
- Read, M., Harrison, R. J., Romagnoli, B., Tanious, F. A., Gowan, S. H., Reszka, A. P., Wilson, W. D., Kelland, L. R., and Neidle, S. (2001). Structure-based design of selective and potent G quadruplex-mediated telomerase inhibitors. *Proc. Natl. Acad. Sci. USA*, **98**, 4844–4849.
- Read, M.A. and Neidle, S. (2000). Structural characterization of a guanine-quadruplex ligand complex. *Biochemistry*, **39**, 13422–13432.
- Riou, J.-F., Guittat, L., Mailliet, P., Laoui, A., Renou, E., Petitgenet, O., Mégnin-Chanet, F., Hélène, C., and Mergny, J.-L. (2002). Cell senescence and telomere shortening induced by a new series of specific G-quadruplex DNA ligands. *Proc. Natl. Acad. Sci. USA*, **99**, 2672–2677.
- Rohs, R., Blockh, I., Sklenar, H., and Shakked, Z. (2005). Molecular flexibility in *ab initio* drug docking to DNA: binding-site and binding-mode transitions in all-atom Monte Carlo simulations. *Nucleic Acids Res.*, **33**, 7048–7057.
- Rueda, M., Cubero, E., Laughton, C. A., and Orozco, M. (2004). Exploring the counterion atmosphere around DNA: what can be learned from molecular dynamics simulations? *Biophys. J.*, **87**, 800–811.
- Saenger, W., Hunter, W. N., and Kennard, O. (1986). DNA conformation is determined by economics in the hydration of phosphate groups. *Nature*, **324**, 385–388.
- Sastry, M., Fiala, R., Lipman, R., Tomasz, M., and Patel, D. J. (1995). Solution structure of the monoalkylated mitomycin C-DNA complex. *J. Mol. Biol.*, **247**, 338–359.
- Schneider, B. and Berman, H. M. (1995). Hydration of the DNA bases is local. *Biophysical J.*, **69**, 2661–2669.
- Schneider, B., Cohen, D., and Berman, H. M. (1992). Hydration of DNA bases: analysis of crystallographic data. *Biopolymers*, **32**, 725–750.
- Schneider, B., Patel, K., and Berman, H. M. (1998). Hydration of the phosphate group in double-helical DNA. *Biophysical J.*, **75**, 2422–2434.
- Seeman, N. C. (2005). Structural DNA nanotechnology: an overview. *Methods Mol. Biol.*, **303**, 143–166.
- Seeman, N. C., Rosenberg, J. M., and Rich, A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl. Acad. Sci. USA*, **73**, 804–808.
- Shui, X., McFail-Isom, L., Hu, G. G., and Williams, L. D. (1998a). The B-DNA dodecamer at high resolution reveals a spine of water on sodium. *Biochemistry*, **37**, 8341–8355.
- Shui, X., Sines, C. C., McFail-Isom, L., Van Derveer, D., and Williams, L. D. (1998b). Structure of the potassium form of CGCGAATTCGCG: DNA deformation by electrostatic collapse around inorganic cations. *Biochemistry*, **37**, 16877–16887.
- Silver, G. C., Sun, J.-S., Nguyen, C. H., Boutorine, A. S., Bisagni, E., and Hélène, C. (1997). Stable triple-helical DNA complexes formed by benzopyridoindole- and benzopyridoquinoxaline- oligonucleotide conjugates. *J. Amer. Chem. Soc.*, **119**, 263–268.
- Silverman, A. P., Bu, W., Cohen, S. M., and Lippard, S. J. (2002). 2.4-Å crystal structure of the asymmetric platinum complex [Pt(amine)(cyclohexylamine)]²⁺ bound to a dodecamer DNA duplex. *J. Biol. Chem.*, **277**, 49743–49749.

- Spingler, B., Whittington, D. A., and Lippard, S. J. (2001). 2.4 Å crystal structure of an oxaliplatin 1,2-d(GpG) intra-strand cross-link in a DNA dodecamer duplex. *Inorg. Chem.*, **40**, 5596–5602.
- Spink, N., Brown, D. G., Skelly, J. V., and Neidle, S. (1994). Sequence-dependent effects in drug-DNA interaction: the crystal structure of Hoechst 33258 bound to the d(CGCAAATTTGCG)₂ duplex. *Nucleic Acids Res.*, **22**, 1607–1612.
- Subramaniam, G., Paz, M. M., Kumar, G. S., Das, A., Palom, Y., Clement, C. C., Patel, D. J., and Tomasz, M. (2001). Solution structure of a guanine-N7-linked complex of the mitomycin C metabolite 2,7-diaminomitosene and DNA. Basis of sequence selectivity. *Biochemistry* **40**, 10473–10484.
- Sun, D., Thompson, B., Cathers, B. E., Salazar, M., Kerwin, S. M., Trent, J. O., Jenkins, T. C., Neidle, S., and Hurley, L. H. (1997). Inhibition of human telomerase by a G-quadruplex-interactive compound. *J. Med. Chem.*, **40**, 2113–2116.
- Takahara, P. M., Frederick, C. A., and Lippard, S. J. (1996). Crystal structure of the anticancer drug cisplatin bound to duplex DNA. *J. Amer. Chem. Soc.*, **118**, 12309–12321.
- Tanious, F. A., Laine, W., Peixoto, P., Bailly, C., Goodwin, K. D., Lewis, M. A., Long, E. C., Georgiadis, M. M., Tidwell, R. R., and Wilson, W. D. (2007). Unusually strong binding to the DNA minor groove by a highly twisted benzimidazole diphenylether: induced fit and bound water. *Biochemistry, in press*
- Temperini, C., Cirilli, M., Aschi, M., and Ughetto, G. (2005). Role of the amino sugar in the DNA binding of disaccharide anthracyclines: crystal structure of the complex MAR70/d(CGATCG). *Bioorg. Med. Chem.*, **13**, 1673–1679.
- Tereshko, V., Minasov, G., and Egli, M. (1999). The Dickerson–Drew B-DNA dodecamer revisited at atomic resolution. *J. Amer. Chem. Soc.*, **121**, 470–471.
- Teuben, J.-M., Bauer, C., Wang, A. H.-J., and Reedijk, J. (1999). Solution structure of a DNA duplex containing a *cis*-diammineplatinum(II) 1,3-d(GTG) intrastrand cross-link, a major adduct in cells treated with the anticancer drug carboplatin. *Biochemistry*, **38**, 12305–12312.
- Thompson, A. S. and Hurley, L. H. (1995). Solution conformation of a bizelesin A-tract duplex adduct: DNA-DNA cross-linking of an A-tract straightens out bent DNA. *J. Mol. Biol.*, **252**, 86–101.
- Thorpe, J. H., Hobbs, J. R., Todd, A. K., Denny, W. A., Charlton, P., and Cardin, C. J. (2000). Topoisomerase poison in the presence of Co²⁺ ions. *Biochemistry*, **39**, 15055–15061.
- Todd, A. K., Adams, A., Thorpe, J. H., Denny, W. A., Wakelin, L. P. G., and Cardin, C. J. (1999). Major groove binding and “DNA-induced” fit in the intercalation of a derivative of the mixed topoisomerase I/II poison N-(2-(dimethylamino)ethyl)acridine-4-carboxamide (DACA) into DNA: X-ray structure complexed to d(CG(5-BrU)ACG)₂ at 1.3-Å resolution. *J. Med. Chem.*, **42**, 536–540.
- Tomasz, M. and Palom, Y. (1997). The mitomycin bioreductive antitumor agents: cross-linking and alkylation of DNA as the molecular basis of their activity. *Pharmacol. Therap.*, **76**, 73–87.
- Van Hecke, K., Nam, P. O. C., Nguyen, M. T., and Van Meervelt, L. (2005). Netropsin interactions in the minor groove of d(GGCCAATTGG) studied by a combination of resolution enhancement and ab initio calculations. *FEBS J.*, **272**, 3531–3541.
- Vega, M. C., García Sáez, I., Aymami, J., Eritja, R., Van der Marel, G. A., Van Boom, J. H., Rich, A., and Coll, M. (1994). Three-dimensional crystal structure of the A-tract DNA dodecamer d(CGCAAATTTGCG) complexed with the minor-groove-binding drug Hoechst 33258. *Eur. J. Biochem.*, **222**, 721–726.
- Viger, A. and Dervan, P. B. (2006). Exploring the limits of benzimidazole DNA-binding oligomers for the hypoxia inducible factor (HIF) site. *Bioorg. Med. Chem.*, **14**, 8539–8549.
- Wang, A. H.-J., Ughetto, G., Quigley, G. J., and Rich, A. (1987). Interactions between an anthracycline antibiotic and DNA: molecular structure of daunomycin complexed to d(CpGpTpApCpG) at 1.2-Å resolution. *Biochemistry*, **26**, 1152–1163.
- Wang, A. H.-J., Ughetto, G., Quigley, G. J., Hakoshima, T., van der Marel, G. A., van Boom, J. H., and Rich, A. (1984). The molecular structure of a DNA-triostin A complex. *Science*, **225**, 1115–1121.
- Wang, L., Bailly, C., Kumar, A., Ding, D., Bajic, M., Boykin, D. W., and Wilson, W. D. (2000). Specific molecular recognition of mixed nucleic acid sequences: an aromatic dication that binds in the DNA minor groove as a dimer. *Proc. Natl. Acad. Sci. USA*, **97**, 12–16.
- Wang, H. and Laughton, C. A. (2007). Molecular modelling methods for prediction of sequence-selectivity in DNA recognition. *Methods*, **42**, 196–203.
- Waring, M. J. (1970). Variation of the supercoils in closed circular DNA by binding of antibiotics and drugs: Evidence for molecular models involving intercalation. *J. Mol. Biol.*, **54**, 247–279.
- Warren, C. L., Kratochvil, N. C. S., Hauschild, K. E., Foister, S., Brezinski, M. L., Dervan, P. B., Phillips, G. N. Jr., and Ansari, A. Z. (2006). Defining the sequence-recognition profile of DNA-binding molecules. *Proc. Natl. Acad. Sci. USA*, **103**, 867–872.
- Wheelhouse, R. T., Sun, D., Han, H., Han F. X., and Hurley, L. H. (1998). Cationic porphyrins as telomerase inhibitors: the interaction of tetra-(N-methyl-4-pyridyl)porphine with quadruplex DNA. *J. Amer. Chem. Soc.*, **120**, 3261–3262.

- Williams, H.E.L. and Searle, M.S. (1999). Structure, dynamics and hydration of the nogalamycin-d(ATGCAT)₂ complex determined by NMR and molecular dynamics simulations in solution. *J. Mol. Biol.*, **290**, 699–716.
- Wilson, W. D., Nguyen, B., Tanious, F. A., Mathis, A., Hall, J. E., Stephens, C. A., and Boykin, D. W. (2005). Dications that target the DNA minor groove: compound design and preparation, DNA interactions, cellular distribution and biological activity. *Curr. Med. Chem. Anticancer Agents*, **5**, 389–408.
- Woods, K. C., Martin, S. S., Chu, V. C., and Baldwin, E. P. (2001). Quasi-equivalence in site-specific recombinase structure and function: crystal structure and activity of trimeric Cre recombinase bound to a three-way Lox DNA junction. *J. Mol. Biol.*, **313**, 49–69.
- Wu, Y., Bhattacharyya, D., King, C. L., Baskerville-Abraham, I., Huh, S. H., Boysen, G., Swenberg, J. A., Temple, B., Campbell, S. L., and Chasney, S. G. (2007). Solution structures of a DNA dodecamer duplex with and without a cisplatin 1,2-d(GG) intrastrand cross-link: comparison with the same DNA duplex containing an oxaliplatin 1,2-d(GG) intrastrand cross-link. *Biochemistry*, **46**, 6477–6487.
- Yang, X.-L., Robinson, H., Gao, Y.-G., and Wang, A. H.-J. (2000). Binding of a macrocyclic bisacridine and ametrone to CGTACG involves similar unusual intercalation platforms. *Biochemistry*, **39**, 10950–10957.
- Zahler, A. M., Williamson, J. R., Cech, T. R., and Prescott, D. M. (1991). Inhibition of telomerase by G-quartet DNA structures. *Nature*, **350**, 718–720.
- Zhang, Q. A., Dwyer, T. J., Tsui, V., Case, D. A., Cho, J., Dervan, P. B., and Wemmer, D. E. (2004). NMR structure of a cyclic polyamide-DNA complex. *J. Amer. Chem. Soc.*, **126**, 7958–7966.
- Zimmer, C. H. and Wähnert, U. (1986). Non-intercalating DNA-binding ligands: specificity of the interaction and their use as tools in biophysical, biochemical and biological investigations of the genetic material. *Prog. Biophys. Mol. Biol.*, **47**, 31–112.

Further Reading

DNA-Water Interactions

- Berman, H. M. (1991). Hydration of DNA. *Curr. Opin. Struct. Biol.*, **1**, 423–427.
- Berman, H. M. (1994). Hydration of DNA: take 2. *Curr. Opin. Struct. Biol.*, **4**, 345–350.
- Westhof, E. (1988). Water: an integral part of nucleic acid structure. *Ann. Rev. Biophys.*, **17**, 125–144.
- Westhof, E. and Beveridge, D. L. (1990). In *Water Science Reviews 5*. (Ed. Franks, F.) Cambridge University Press, Cambridge.

DNA-Drug Interactions

- Denison, C. and Kodadek, T. (1998). Small-molecule-based strategies for controlling gene expression. *Chem. Biol.*, **5**, 129–145.
- Dervan, P. B. (2001). Molecular recognition of DNA by small molecules. *Bioorg. Med. Chem.*, **9**, 2215–2235.
- Dervan, P. B. and Edelson, B. S. (2003). Recognition of the DNA minor groove by pyrrole-imidazole polyamides. *Curr. Opin. Struct. Biol.*, **13**, 284–289.
- Gale, E. F., Cundliffe, E. F., Reynolds, P. E., Richmond, M. H., and Waring, M. J. (1981). *The Molecular Basis of Antibiotic Action*. John Wiley, London.
- Gao, X. and Patel, D. J. (1989). Antitumour drug-DNA interactions: NMR studies of echinomycin and chromomycin complexes. *Quart. Rev. Biophys.*, **22**, 93–138.
- Hurley, L. H. (2002). DNA and its associated processes as targets for cancer therapy. *Nature Rev. Cancer*, **2**, 188–200.
- Neidle, S. (2001). DNA minor-groove recognition by small molecules. *Natural Product Rep.*, **18**, 291–309.
- Neidle, S. and Waring, M. J. Eds., (1993). *Molecular Basis of Drug-DNA Antitumour Action*. Macmillan Press, London. Vols 1, 2.
- Pratt, W. B., Ruddon, R. W., Enslinger, W. D., and Maybaum, J. (1994). *The Anticancer Drugs, 2nd Ed.*, Oxford University Press, New York.
- Rajski, S. R. and Williams, R. M. (1998). DNA cross-linking agents as antitumor drugs. *Chem. Rev.*, **98**, 2723–2796.
- Reddy, B. S. P., Sondhi, S. M., and Lown, J. W. (1999). Synthetic DNA minor groove-binding drugs. *Pharmacol. Therap.*, **84**, 1–111.

- Ren, J. and Chaires, J. B. (1999). Sequence and structural selectivity of nucleic acid binding ligands. *Biochemistry*, **38**, 16067–16075.
- Thurston, D. E. (1999). Nucleic acid targeting: therapeutic strategies for the 21st century. *Brit. J. Cancer*, **80**, 65–85.
- Wemmer, D. E. and Dervan, P. B. (1997). Targeting the minor groove of DNA. *Curr. Opin. Struct. Biol.*, **7**, 355–361.
- Yang, X.-L. and Wang, A. H.-J. (1999). Structural studies of atom-specific anticancer drugs acting on DNA. *Pharmacol. Therap.*, **83**, 181–215.

RNA Structures and Their Diversity

6.1 Introduction

The preceding chapters have illustrated the structural variability displayed by DNA. Virtually all of this is within the boundaries imposed by the overriding structural requirements of double-, triple-, or four-stranded helices. RNA molecules are not subject to these limitations, and consequently their structural variety is in striking contrast to the relative predictability of most DNA tertiary structure (though the increasing range of topologies shown by quadruplex DNAs suggests that DNA can be structurally highly versatile under appropriate circumstances when released from the constraints of the double helix). It is perhaps unsurprising (at least in retrospect!) that the ability of many RNAs to fold into compact tertiary structures, analogous to the folding of polypeptides into proteins and enzymes, is paralleled in some instances, by the possession of catalytic activity.

The need for an RNA intermediary between the code for a gene contained in DNA, and the ultimate gene product, a protein, was realized soon after the discovery of the structure of DNA itself. This category of (single-stranded) RNA molecule was initially conceived of as a carrier of the code to the ribosome, although this picture of RNA as an intermediary is somewhat complicated by the possession of double-stranded RNA by RNA viruses. The need for a second type of RNA molecule, transfer RNA (tRNA), was also recognized early in the history of molecular biology, in order to perform the specialized function of decoding and transferring the information contained in an individual codon in a messenger RNA molecule into the correct amino acid. An organism has 20 sets of tRNA molecules, each set specific for a particular amino acid (in principle only one tRNA is required for each amino acid, but in practice there are more than this, with some redundancy being needed for several amino acids). A specific tRNA attaches a correct amino acid and transfers it to a specialized region of the ribosome, where it is incorporated into a growing protein chain. A third yet more specialized function for RNA has been found, that of RNA enzymatic activity, for catalytic cleavage of RNA (and DNA) itself. Such cleavage can be intramolecular, such as self-splicing introns in RNA editing. RNA cleavage by RNA molecules, termed ribozymes, has excited much interest with the suggestion that they could be potential key players in a prebiotic pre-DNA world. Ribozymes are also being studied for their potential therapeutic use in gene therapy,

for cleavage and ultimate destruction of specific DNA sequences. A further function arises from a more recent discovery, of some complex RNA sequences that can fold and act as molecular switches (riboswitches) on binding certain small molecules. Finally, and most surprisingly, has been the far-reaching discovery that short 20–25 nucleotide long double-stranded RNAs (small interfering RNAs, or siRNAs), can inhibit eukaryotic gene expression. They do this via the RNA interference (RNAi) pathway.

RNA crystallographic and NMR structural studies have advanced rapidly over the past few years. There have been several technical advances that have made this possible. On the one hand, solid-phase RNA chemical synthesis and purification (of sequences up to 15–20 bases in length) is now usually relatively straightforward for the preparation of milligram quantities, although the methods are still not as routine as those for DNA oligomer preparation (Kundrot, 1997; Anderson *et al.*, 1996). Even short RNA sequences are very prone to secondary structure formation, and especial care in handling and purification is needed. Larger sequences are routinely prepared enzymatically, using efficient *in vitro* transcription of the appropriate DNA sequence with T7 RNA polymerase, although there is a requirement for a purine sequence at the 5' end in order for optimal efficiency of transcription to take place. Also, random sequence nucleotides are sometimes added at the 3' end by T7 polymerase. These problems can be overcome by methods that can posttranscriptionally excise the desired sequence (Price *et al.*, 1995). Commercial kits are widely available for RNA transcription, which can readily produce milligram quantities of sequences up to several hundred bases in length.

Crystallization of RNAs has traditionally been difficult, in part on account of the inherent flexibility of many RNA molecules. The introduction of sparse matrix methods coupled with robotic technology has enabled success to be achieved relatively rapidly with a number of RNAs (Doudna *et al.*, 1993; Dock-Bregeon, Moras, and Giegé, 1999; Cate and Doudna, 2000; Ke and Doudna, 2004; Golden, 2007). These methods enable a very large range of crystallization conditions to be rapidly sampled, using only small quantities of RNA. However success by screening alone is never guaranteed, and methods have been successfully developed which rationally engineer specific intermolecular contacts in order to optimize the formation of crystalline arrays of molecules (Ferre-D'Amare, Zhou, and Doudna, 1998). Crystal structures of the majority of small, purely helical RNAs are still routinely solved by molecular replacement methods, normally using canonical duplex RNA models as starting points. Synthetic RNAs, as with DNAs, can incorporate heavy atoms such as bromine for MAD phasing, on specific residues. These are needed when molecular replacement does not work, or with more complex structures (see, e.g., Wedekind and McKay, 2000). Large, complex RNAs always require heavy-atom derivatives, and systematic approaches have been developed for rapidly homing in on those heavy atom reagents that are most likely to produce useable derivatives, such as osmium hexamine. A useful strategy in some instances is to use a small synthetic RNA sequence containing a heavy-atom derivative, which can become an integral part of the total RNA molecule when annealed with the larger RNA component prepared by transcription. The incorporation of selenium in 2'-methylseleno derivatives, synthesized either chemically or enzymatically, is of particular use for MAD phasing (Höbartner *et al.*, 2005).

6.2 Fundamentals of RNA Structure

6.2.1 Helical RNA Conformations

RNA differs from DNA in two key chemical aspects (Fig. 6.1):

- The sugar groups in all RNAs have an OH group attached at the 2' position, making them ribose rather than deoxyribose sugars
- Thymine bases are replaced by uracils, which have the same hydrogen-bonding, but do not have a methyl group at the 5' position

The effect of the extra hydroxyl group is profound. RNA sugars are much more rigid than DNA ones, and only adopt the *C3'-endo* pucker type (Fig. 6.2) since otherwise the 2'-OH group would clash with C8 (for a purine) or C6 (for a pyrimidine) of the attached base. Simulation studies, together with analysis of crystal structures, have shown that this pucker is stabilized in several positive ways (Auffinger and Westhof, 1997).

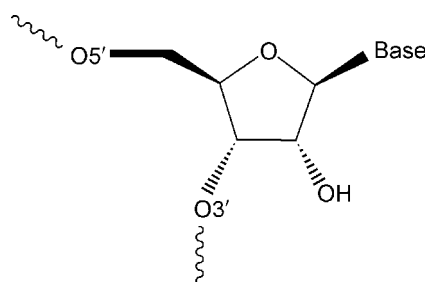


Figure 6.1 Configuration of a ribose sugar in RNA.

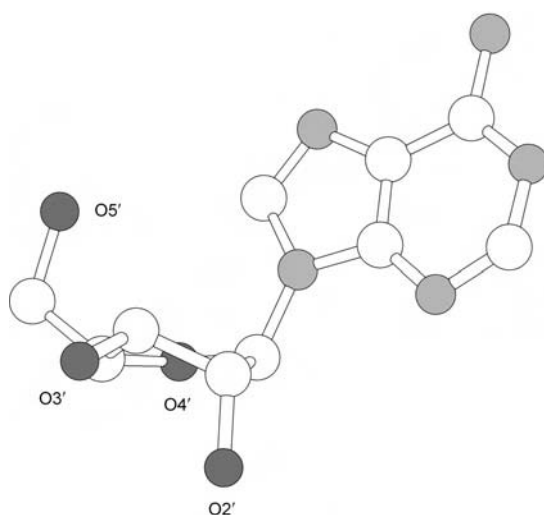


Figure 6.2 A ribonucleoside in an A-RNA conformation.

The 2'-OH group is oriented only in several discrete positions as a result of interactions with O3' or O4', or with accessible hydrogen-bond acceptors on the attached base. In helical A-RNA, the important interaction is between the H2' atom and O4' of the next residue, which is not accessible for hydrogen-bonding to water molecules.

Single-stranded RNAs do not adopt rigid conformations in the absence of restraining tertiary interactions. However appropriate sequences readily form antiparallel double-stranded RNA structures, analogous to duplex DNA. Uracil participates in U•A base pairs that are fully isomorphous with A•T ones in duplex DNA. Duplex RNA is conformationally rather rigid, and its behavior contrasts remarkably with the polymorphism of duplex DNA in that only one major polymorph of the RNA double helix has been observed. This has many features in common with A-DNA, and accordingly is known as A-RNA. The conformational features of canonical RNA helices have been obtained from fiber-diffraction studies on duplex RNA polynucleotides from both viral and synthetic origins. A-RNA is an 11-fold helix, with a narrow and deep major groove and a wide, shallow minor groove, and base pairs inclined to and displaced from the helix axis (Fig. 6.3a,b). It is notable that the base pairs themselves in canonical A-RNA are only slightly twisted. The major A-RNA form is apparent in natural polynucleotide fibres, or in those formed from poly(A)•poly(T) in conditions of low ionic strength (Arnott *et al.*, 1973). At higher ionic strength, a slightly different 12-fold helix (A'-RNA) is formed, which retains the same overall helical features, although its wider major groove resembles that of A-DNA. These are detailed in Table 6.1. Both helices have the C3'-*endo* sugar pucker characteristic of ribopolynucleotides. Another difference from duplex DNA is that RNA helices, though capable of a small degree of bending (up to ~15°), does not undergo the large-scale bending seen, for example, in A-tract DNA.

A large number of single-crystal and NMR studies on sequences forming perfectly base-paired duplex RNA have been performed, which have shown that the RNA geometry observed in fibers is conserved across a wide range of sequences. This is provided that there are no base mismatches, which can under some circumstances, greatly alter helical structure (see below). As a consequence of their greater conformational rigidity RNA helices do not have the pronounced sequence-dependent microstructure of duplex DNA. This has received confirmation from several database surveys, including one of 3000 nucleotides from ribosomal crystal structures (Schneider, Morávek, and Berman, 2004), which found that most are A-type (i.e. helical). The 18 other conformational types are representative of the various nonhelical motifs that are characteristic of complex RNA molecules, and which form much of the subject matter of this chapter.

The first reported single-crystal duplex RNA analyses were of the self-complementary dinucleosides r(AU) and r(GC) (Seeman *et al.*, 1976; Rosenberg *et al.*, 1976). These determinations, at atomic resolution, are of historic importance in showing Watson–Crick base pairing within the context of short helical segments, and in demonstrating that even dinucleoside duplexes have backbone conformations characteristic of helical A-RNA. They also provided the impetus for structural studies of intercalative drug-binding to dinucleoside duplexes as models for the polynucleotide situation (see Chap. 5).

Crystallographic analyses of sequences such as the octanucleotide r(CCCCGGGG), have shown helices of length sufficient for a full set of helical parameters to be extracted. This sequence crystallises in two distinct crystal lattices, enabling the effects of crystal packing factors on structure to be assessed (Portmann, Usman, and Egli, 1995). In each instance, rhombohedral and hexagonal, the RNA helices are very similar, and their features closely resemble those in fiber-diffraction canonical

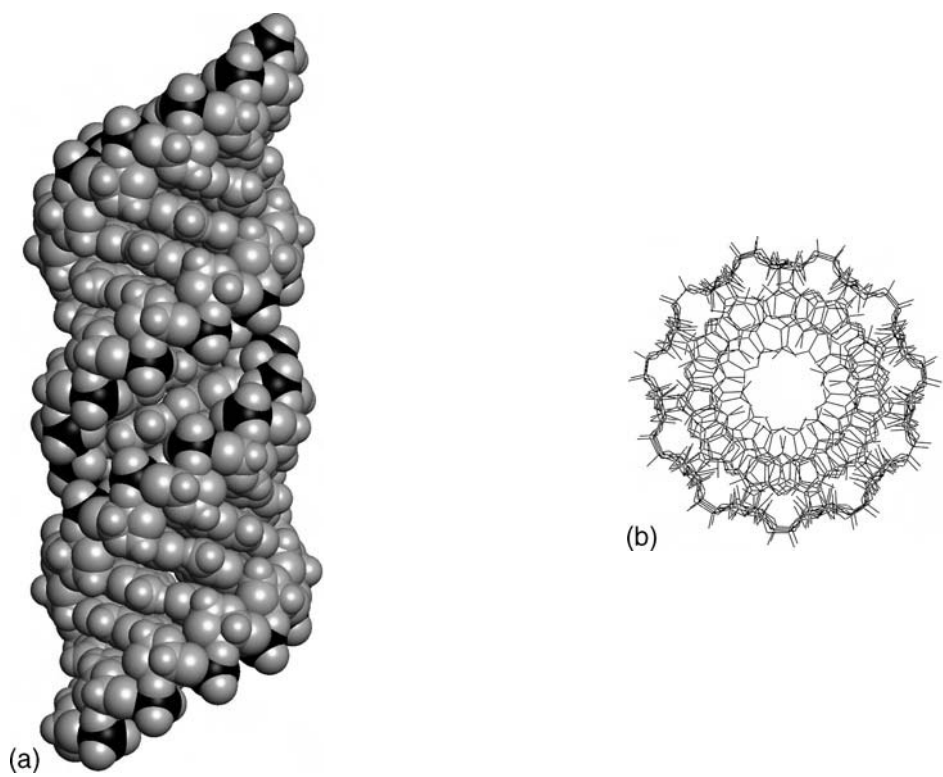


Figure 6.3 (a) Space-filling representation of an A-RNA double helix. (b) View down the helix axis of an A-RNA double helix, drawn in stick mode. Note that the base pairs sit astride the helix axis.

Table 6.1 Structural Features of RNA Double Helices, from Fiber Diffraction Studies (Arnott, 1999)

(a) Conformational angles. Values for poly(A)•poly(T) in the A-form are given

		α	β	χ	δ	ϵ	ξ	ζ
A-RNA	Adenine	-69	179	55	82	-154	-71	-161
	Thymine	-64	178	51	83	-152	-173	-161
A'-RNA		-70	177	61	77	-153	-70	-163

(b) Groove dimensions in Å

	Major-groove width	Major-groove depth	Minor-groove width	Minor-groove depth
A-RNA	4.7	12.9	10.8	3.3
A'-RNA	8.9	14.4	10.5	3.4

(c) Selected helical and base/base pair morphological parameters

	Base displacement (Å)	Rise (Å)	Helical twist (°)	Roll (°)	Propeller twist (°)	Inclination (°)
A-RNA	0.44	2.8	32.7	-0.8	-2.1	-15.5
A'-RNA	0.51	3.0	30.0	0.0	2.3	-10.6

A-RNA with, for example, minor groove widths of 9.8 Å and 9.7 Å respectively. Visualization of a complete turn of RNA double helix requires a longer sequence. The structure of r(UUAUAUAUAUAUA) was the first to show this (Dock-Bregeon *et al.*, 1989). Again, the helix is essentially classical A-RNA (Fig. 6.4), with an average helical twist of 33° and an axial rise of 2.8 Å. The structure of a mixed sequence, r(UAAGGAGGUGAU), has been determined (Schindelin *et al.*, 1995), which is a more rigorous test of the conservation of RNA helical structure. This RNA dodecamer sequence closely resembles a region required for the initiation of translation in prokaryotes. The overall features of the two independent helices in this crystal structure are of A-RNA type rather than A'-RNA with, for example, helical twists of 33.3° and 33.5° respectively, corresponding to a helical repeat of 10.7–10.8 base pairs/turn.

Longer RNA duplex sequences are more difficult to crystallize in ordered conformations. This is illustrated by the analysis of two duplexes, an 18- and a 19-mer, which diffract to high resolution, 1.20 Å and 1.55 Å respectively (Klosterman, Shah, and Steitz, 1999). In spite of this apparent indication of a high degree of internal order, the 18-mer is disordered such that each base pair is an average of the duplex as a whole. This is closely analogous to the fiber diffraction situation, and occurs in both long DNA and RNA duplexes when the crystal-packing involves continuous helices. The 19-mer is disordered in a distinct way, with two orientations being apparent. Both sequences are in classic A-RNA conformations, except for the sugar puckers, which have significantly larger pucker amplitudes than in simple ribonucleosides. Analogous static disorder has been observed in the high-resolution (1.58 Å) crystal structure (Mooers, Logue, and Berglund, 2005) of the 18-mer duplex with the sequence r(CUG)₆, which is implicated in the genetic disease myotonic dystrophy. The structure contains two lattices related by a translation of one helical turn, with each double helix packing head to tail in a pseudo-continuous



Figure 6.4 The crystal structure of the duplex of r(UUAUAUAUAUAUA), shown in schematic form (Dock-Bregeon *et al.*, 1989). Note the characteristic A-RNA inclination of base pairs with respect to the vertical helix axis.

manner. The helix is in an undistorted A-form structure, even though it contains six non-hydrogen-bonded U•U base pairs, which alter the pattern of electrostatic potential in the minor groove, to one that may be important for binding to specific proteins implicated in this disease.

The availability of high-resolution crystal structures has enabled RNA hydration in oligonucleotides to be defined in detail (Sundaralingam and Pan, 2002). Some general rules have emerged for the water arrangements beyond the obvious finding of an inherently greater degree of hydration compared to DNA oligonucleotides, on account of the 2'-OH group being an active and willing hydrogen-bond participant. There is no analogue of the minor-groove spine of hydration seen in B-DNA. Again this is unsurprisingly since in A-RNA the minor groove is too wide for such an arrangement. A detailed analysis (Egli, Portmann, and Usman, 1996) of the crystal structure of r(CCCCGGGG) at a resolution of 1.8 Å has shown that the base edges and phosphate groups tend to be more extensively hydrated than in DNA. This appears to be in large part due to the 2'-OH groups, with the 16 in this duplex having 33 hydrogen bonds to water molecules. These 2'-OH act as foci for the establishment of water networks in both major and minor grooves. There is, for example, a set of seven identical well-defined transverse hydrogen-bonded water linkages across the minor groove (Fig. 6.5). Long time-scale molecular dynamics simulations on equivalent DNA and RNA oligomers have confirmed the slightly greater hydration of the latter (Auffinger and Westhof, 2000), with about 21.4 water molecules per RNA C•G base pair compared to 19.8 for a B-DNA one. Counter-ions such as sodium or potassium tend to accumulate in the major groove of RNA duplexes (Auffinger and Westhof, 2000; Cheatham and Kollman, 1997), by contrast with their minor-groove localization in B-DNA.

6.2.2 Mismatched and Bulged RNA Structures

RNAs can readily form stable base-pair and triplet mismatches and extrahelical regions within the context of “normal” RNA double helices. Such features, notably the extrahelical loops (Fig. 6.6), have been the subject of intense structural study, since they are present in all complex RNAs (tRNAs, mRNAs, ribozymes, riboswitches, ribosomal RNA), and together with various types of base-stacking, are responsible for maintaining their tertiary structures. It is paradoxically common for crystal structures of short sequences containing potential loop regions such as the UUCG “tetraloop” (which forms an especially stable extrahelical loop structure in solution), to often not show such features. This may be a consequence of the high ionic strength of many crystallization conditions, together with the preference of some sequences to pack in the crystal as helical arrays. So, instead of loops, these crystal structures tend to have runs

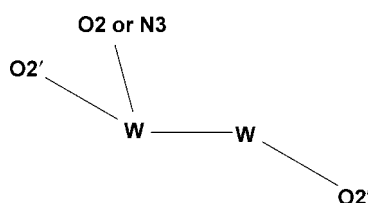


Figure 6.5 The transverse hydrogen-bonded water linkages across the minor groove in the structure of r(CCCCGGGG) (Egli, Portmann, and Usman, 1996).

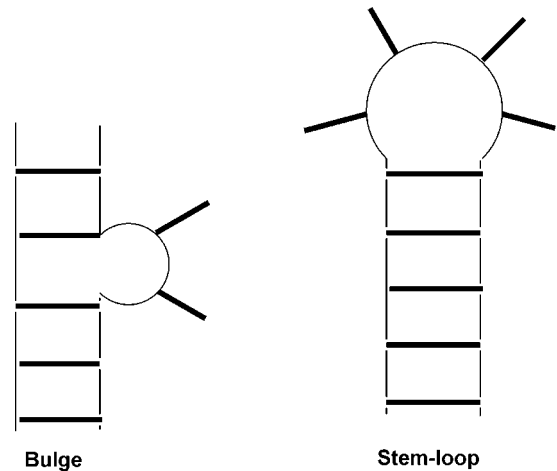


Figure 6.6 Two types of extrahelical features in RNA. The thick lines represent base pairs in duplex regions, and unpaired bases in the loops.

of non-Watson–Crick mismatched base pairs where the loop would have formed. For example, an A-RNA double helix, albeit with G•U and U•U base pairs, is formed in the crystal (Cruse *et al.*, 1994) by the sequence r(GCUUCGGC)d^(Br)U). There is evidence of some deviations from the exact canonical A-RNA duplexes formed by fully Watson–Crick base pairs, since this helix has <10 base pairs per turn. The dodecamer sequence r(GGACUUCGGUCC) similarly forms a base-paired duplex (Holbrook *et al.*, 1991), with U•C and G•U base pairs. The A-RNA helix here has a significant increase in the width of its major groove, possibly on account of the water molecules that are strongly associated with the mismatch base pairs.

Much greater perturbations are apparent (Baeyens *et al.*, 1996) in the structure of the dodecamer r(GGCCGAAAGGCC), where the four non-Watson–Crick base pairs in the center of the sequence form an internal loop with sheared G•A and A•A base pairs. The resulting structure is highly distorted from A-RNA ideality (Fig. 6.7), with a compression of the major groove, an enlargement of the minor groove width to 13.5 Å, and a pronounced curvature of the resulting helix. Again, this sequence forms a tetraloop structure in solution (Heus and Pardi, 1991). The G•U base pair is a prominent and very important element of large RNA structures since it is especially stable (Varani and McClain, 2000), on account of its two hydrogen bonds (Fig. 6.8). It is known as the “wobble” pair, since a G in the first position of a codon can accept either a C or a U in the third anticodon (wobble) position.

Table 6.2 Selected RNA Helix Crystal Structures

Sequence	PDB code no	NDB code no
r(CCCCGGGG)	259D	ARH074
r(UAAGGAGGUGAU)	1SRD	ARL062
r(UUAUAUAUAUAUA)	1RNA	ARN035
r(GGGGGGGGGGGG)	1QCU	AR0020
r(CUGCUGCUGCUGCUGCUG)	1ZEV	AR0062



Figure 6.7 The structure of the dodecamer r(GGCCGAAAGGCC) (Baeyens *et al.*, 1996).

The RNA genome of the HIV-1 retrovirus contains many nonhelical features, some of which have been studied by structural methods. Its dimerization initiation site has features that act as signals for RNA-packaging. Again the expected secondary structure (containing two loops in a “kissing-loop” arrangement) has not always been observed. Instead, the duplex contains two A•G base pairs (Fig. 6.9), each adjacent to an extra-helical, bulged-out adenosine (Ennifar *et al.*, 1999). A true kissing complex for this RNA has been observed (Ennifar *et al.*, 2001), which shows loop–loop interactions (Fig. 6.10). This involves looped-out bases from one strand forming a run of extra-helical A•U and G•C base pairs with the looped-out bases from the adjacent strand (Fig. 6.11). The tendency of short RNA sequences to maximize helical features is also apparent in the crystal structure of a 29-nucleotide fragment from the signal recognition particle (Wild *et al.*, 1999), which forms 28-mer heteroduplexes rather than a hairpin structure. Even so, this duplex has features of wider interest, since it has a number of non-Watson–Crick base pairs such as a 5'-GGAG/3'-GGAG purine bulge. Their overall effect is to produce backbone distortions so that the helix has non-A-RNA features such as a widening of the major groove, by ~ 9 Å, and local undertwisting of base pairs adjacent to A•C and G•U base pairs.

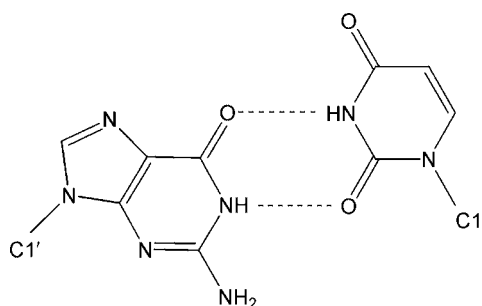


Figure 6.8 The G•U wobble base pair.

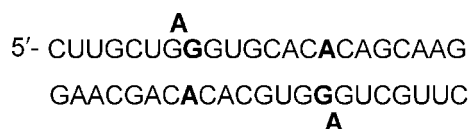


Figure 6.9 The sequence of the dimerization initiation site of the RNA genome from the HIV-1 retrovirus, studied crystallographically (Ennifar *et al.*, 1999).



Figure 6.10 The structure of the coaxially stacked kissing complex of the HIV-1 RNA dimerization initiation site (Ennifar *et al.*, 2001) showing the loop-loop interaction region.

NMR approaches to determining RNA structure have been fruitful in defining the conformation of double-helical RNA. However sometimes, as a consequence of their inherent flexibility, the features of nonhelical regions may be less well-defined. This can lead to a lack of sufficient number of constraints so that these features of an NMR model can show several distinct conformations – such flexible features may well be functionally important. The correspondence (and complementarity) between crystallographic and NMR structures is well-illustrated in studies of the high-affinity RNA binding site of the HIV-1 Rev protein response element, whose conformation changes significantly with bound protein. The NMR and X-ray structures on the

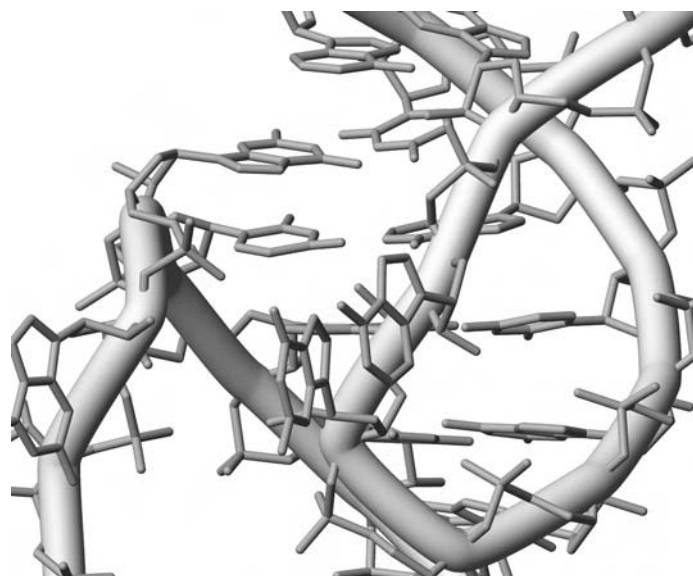


Figure 6.11 Close-up view of the kissing complex.

Table 6.3 Crystal and NMR Structures of Some Small RNAs with Nonhelical Features

Sequence	Feature	PDB code no	NDB code no
r(GGACUUCGUCC)	U•C, G•U base pairs	255D	ARL037
r(GGCCGAAAGGCC)	G•A, A•A base pairs	283D	URL051
29-nucleotide SRP	purine bulge	1D4R	AR0024
r(GCGCGGCACCGUCCGCGGAACAAACGG)	pseudoknot	437D	UR0004
HIV-1 RRE site	G•G, A•G base pairs; A, U extrahelical	1CSL	AR0023
Telomerase RNA domain	pseudoknot	1YMO	–
Viral ribosomal frameshift RNA	pseudoknot	1L2X	UR0020
HIV-1 RNA dimerization	Kissing complex	1K9W	UR0017

RNA itself agree in describing the RRE as a highly distorted helix (Peterson and Feigon, 1996; Ippolito and Steitz, 2000). The distortions are a consequence of (i) two extrahelical bases, an adenine and a uracil from one strand, which are responsible for a bulge and a kink respectively in the helix, and (ii) two essential purine–purine base pairs, which flank the extrahelical uracil (Fig. 6.12). Both structures show that the helix is bent by $\sim 30^\circ$, with a narrowed major groove, and is highly distorted from canonical A-RNA. The principal difference between them is in the loop region around the extrahelical uracil. This is a consequence of the adjacent G(*syn*)•G(*anti*) base pair having an asymmetric configuration in the X-ray structure, whereas in the NMR structure it has been assigned a symmetric one, so that the backbones are forced to adopt distinct paths in this region. Whether these differences are real, or reflect differing solution conditions, remains to be established.

We have noted above that crystallographic studies of small RNAs containing putative loop regions very frequently result in helices that instead maximize base-pairing. The details of the geometry of the structurally important and very stable UUCG tetraloop were especially elusive for some while. Initial NMR studies proved to be only partially correct. One NMR analysis (Allain and Varani, 1995) is in close agreement with a subsequent crystallographic analysis of a tetraloop incorporated in a large RNA-protein crystal structure (Ennifar *et al.*, 2000). Molecular dynamics simulations (Miller and Kollman, 1997) of the tetraloop have converged to the correct structure from an incorrect starting-point. The essential features of the UUCG tetraloop (Fig. 6.13a) are (i) an unusual G(*syn*)•U(*anti*) base pair (Fig. 6.13b), which is seen directly in the X-ray structure, (ii) the 2'-OH group of this uridine nucleoside hydrogen-bonds to the O6 atom of the guanine,



Figure 6.12 The structure of the RNA binding site of the HIV-1 Rev protein response element (Peterson and Feigon, 1996; Ippolito and Steitz, 2000).

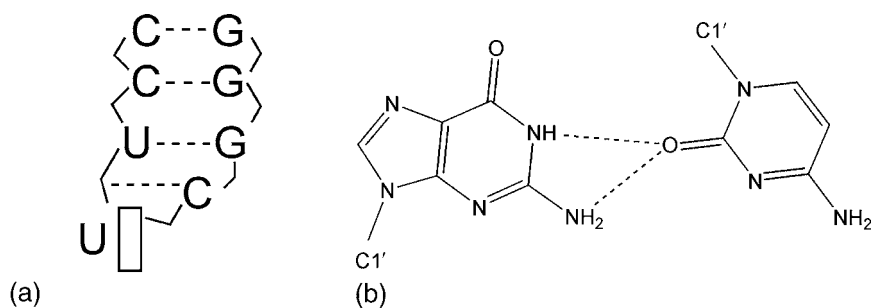


Figure 6.13 (a) A schematic view of the UUCG tetraloop, with the U base looped out from pairing. (b) The G(*syn*)•U(*anti*) base pair found in the UUCG tetraloop (Allain and Varani, 1995; Ennifar *et al.*, 2000; Miller and Kollman, 1997).

(iii) the cytosine stacks on the adjacent uracil base. Tetraloops, and other other elusive motifs, can be captured when they interact with their biological partners, proteins, or other nucleic acids. Thus the solution structure of the AGAA tetraloop, a member of the AGNN family, has been determined bound to the duplex RNA-binding domain of the Rnt1 protein (Wu *et al.*, 2004). Surprisingly, there are no specific hydrogen bonds between the protein and the conserved A or G bases. Instead, recognition is governed by shape complementarity, with an α -helix fitting into the minor groove of the tetraloop.

A more complex motif in RNA is the pseudoknot, which is a prominent feature of many larger RNA structures (see Sect. 6.4 on ribozymes). The detailed structures of a number of pseudoknots have been determined by NMR methods, for example, the telomerase RNA pseudoknot domain (Theimer, Blois, and Feigon, 2005). This structure has a long triple helix close to the highly conserved pseudoknot junction (Fig. 6.14), involving a number of Hoogsteen base triplets (Fig. 6.15). Rather fewer crystallographic analyses have been reported for pseudoknot-containing oligoribonucleotides. Notable studies have included the high-resolution analyses of the pseudoknots from potato leaf roll virus (Pallan *et al.*, 2005) and beet western yellow virus (Egli *et al.*, 2002), at 1.35 Å and 1.25 Å respectively. These have enabled a high level of detail in the structures to be visualized, especially the roles played by water molecules. For instance, there are several instances of water-base “stacking” (Sarkhel, Rich, and Egli, 2003), which involve water molecules being close to the face of an individual base such that water- π orbital interactions are likely to occur.



Figure 6.14 The NMR structure found for the telomerase RNA pseudoknot domain (Theimer, Blois, and Feigon, 2005).

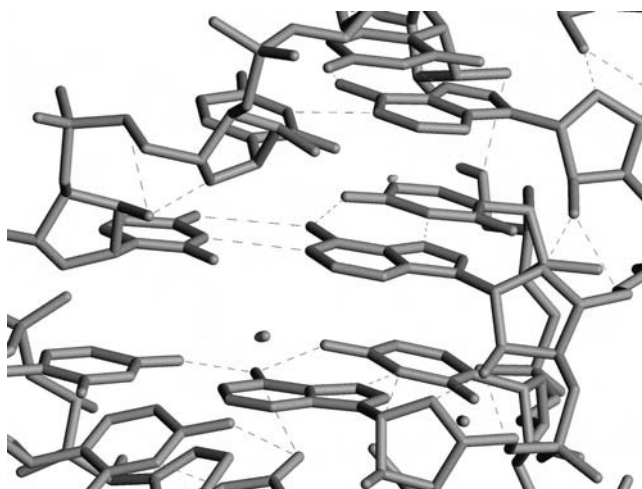


Figure 6.15 Detailed view of the telomerase RNA pseudoknot domain, showing the run of Hoogsteen base triplets.

The relative precision of crystallography and NMR is always hard to assess unless identical structures are compared. This has been done for two small RNAs, one of which is fully helical, and the other has a small loop in it (Rife *et al.*, 1999). Grati-fyingly, the differences between each pair of structures are small. The study also compared various parameter sets commonly used for nucleic acid X-ray and NMR refinements, and concluded that deficiencies in them (especially in electrostatics) are probably responsible for many of the (small) differences observed between solution and crystal structures. The overall message is that it should be the goal of solution structure determination to reproduce as closely as possible the local and global structures found by high-resolution crystallography.

These relatively small RNAs have classic Watson–Crick base pairs in their heli-cal stems. Nonhelical parts show a variety of base–base interactions, which become more important with increased size and complexity of RNAs, as will be discussed in subsequent sections. In principle the total number of non-Watson–Crick base-pair possibilities is very large, as shown in Fig. 6.16, illustrating some of the pos-sible G•A arrangements. This number is considerably increased when sugar group substituents are involved. Even more base triplet possibilities can be envisaged, and very many have been observed, such as the stable G•G•A arrangement (Fig. 6.17). The large number of RNA-pairings that have been observed can be systematically classified according to the relative orientations of the attached nucleotide strand, and the base edges involved (Leontis and Westhof, 2001; Leontis, Stombaugh, and Westhof, 2002).

6.3 Transfer RNA Structures

We have already seen in Chap. 4 that DNA sequences are able to participate in non-Watson–Crick base–base interactions. RNA sequences are inherently even more capa-ble of doing so, and indeed they are the norm in large natural RNA structures, once

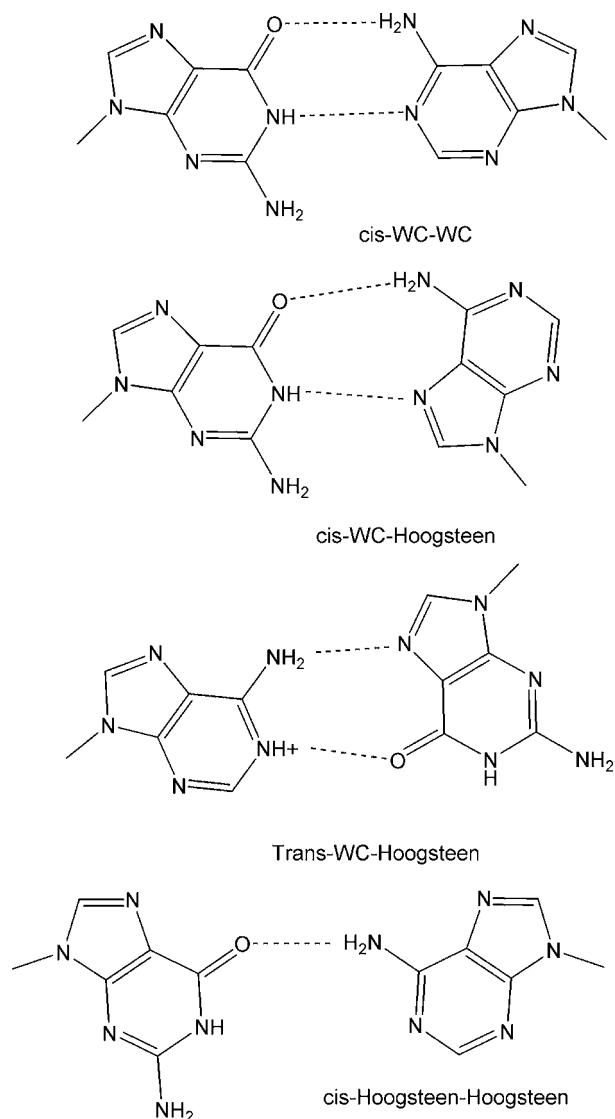


Figure 6.16 Some of the possible G•A base-pair arrangements that have been found in RNA.

Watson–Crick base-pairing is maximized. There is a significant literature on RNA secondary structure prediction, and a number of computer algorithms are freely available with which one can examine RNA sequences. They are based in large part on the maximizing of base-pairing. Successful prediction also takes into account the existence of a number of invariant features in a sequence across species. Historically, the determination of the primary sequence of a tRNA molecule in 1964 rapidly led to suggestions for its secondary folding into a cloverleaf structure. This fold is a consequence of base–base hydrogen-bonding between nonadjacent regions of RNA double helix and loop. It is based on the identification of the highly conserved features of acceptor, anticodon, D and T arms, and three/four loops, one of which is very variable between tRNAs (the

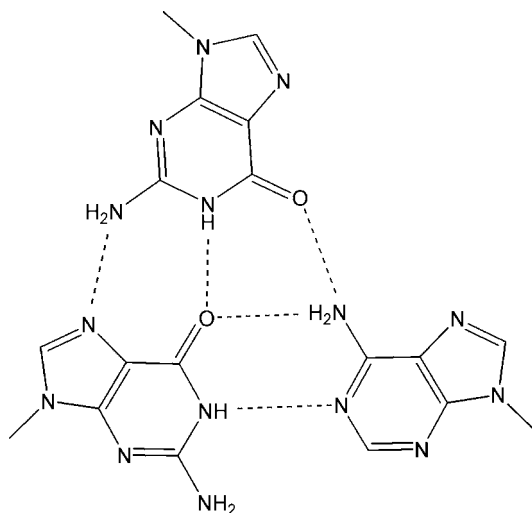


Figure 6.17 A stable G•G•A triplet arrangement.

“variable” loop). There were numerous attempts at predicting the three-dimensional structure of a tRNA. None were completely accurate, which is unsurprising given the very large number of potential base–base interactions. However the general notion that these base–base interactions are responsible for RNA folding, has proved to be correct, as were the assignments of Watson–Crick base pairs. All tRNAs contain, in addition to the standard A, U, G, and C nucleosides, a number of modified nucleosides such as N2-dimethylguanosine, N1-methyladenosine, pseudouridine (Ψ ; an isomer of uridine with the glycosidic bond from the ribose sugar going to the C5 atom of the base rather than to N1), and dihydrouridine D. tRNA species have between 70 and 95 nucleosides, depending in large part on the size of the variable loop.

The crystal structure of yeast phenylalanine tRNA was determined in the early 1970s, simultaneously by two groups (Robertus *et al.*, 1974; Kim *et al.*, 1974). These together with a few other tRNAs (Table 6.4) remained the sole complex RNA molecular structures available for 20 years, until the first ribozyme structure. Both independent structures, one monoclinic (Robertus *et al.*, 1974) and the other orthorhombic (Kim *et al.*, 1974), are closely similar. They showed that the molecule is folded into an overall L-shape, with the two arms at right angles to each other. The original cloverleaf is still apparent, but with additional interactions between distant parts of the structure (Fig. 6.18). The arms consist of short A-RNA helices, together with these extensive base–base interactions that hold the two arms together (Fig. 6.19). The longer double helical anticodon arm has the short helix of the D stem stacked upon it. The other arm is formed by the helix of the acceptor stem, on which is stacked the four base-pair

Table 6.4 Selected tRNA Crystal Structures

tRNA	PDB code no	NDB code no
Yeast Phe, high resolution, monoclinic, 2.0 Å	1EW	TR002
Yeast Phe, high resolution, monoclinic, 1.93 Å	1EHZ	TR0001
Yeast Phe, orthorhombic	6TNA	TRNA04
Yeast Asp	2TRA	TRNA07
Yeast initiator	1YFG	TRNA12

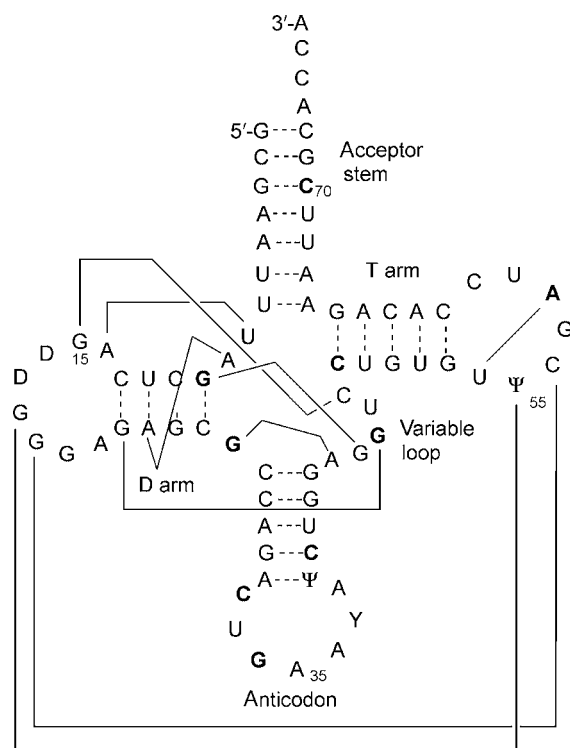


Figure 6.18 Schematic of the hydrogen bonding interactions in the yeast tRNA^{phe} crystal structure (Robertus *et al.*, 1974; Kim *et al.*, 1974). The interactions between nucleotides that appear to be remote in this, the clover-leaf representation.

T arm helix. This key feature of helix–helix stacking has turned out to be of central importance for other complex RNAs. The nine additional base–base interactions that maintain the structural fold all tend to be in the elbow region, where the two arms are joined together. Almost all of these are of non-Watson–Crick type, and several are triplet interactions (Fig. 6.20). These are analogous to the triplets found in DNA triple helices (Chap. 4). Other subsequent crystal structures of tRNAs (Table 6.4) have shown that the overall L shape is invariant, as are many of the tertiary interactions.

The yeast tRNA^{phe} crystal structures have been rerefined several times subsequent to their initial determinations, each time using increasingly robust refinement methods, so that in effect these structures have been used as a test-bed for the general development of nucleic acid crystal structure refinement methodology (Westhof and Sundaralingham, 1986; Westhof, Dumas and Moras, 1988). Recently the crystal structure of the monoclinic form has been redetermined (Shi and Moore, 2000; Jovine, Djordjevic, and Rhodes, 2000) using synchrotron sources to collect data, to 1.93–2.0 Å resolution respectively, from flash-frozen crystals. It is reassuring that the overall backbone geometry and fold remains as found in the earlier studies. There are corrections to the sugar pucker and conformation for a very few nucleotides, and backbone torsion angles conform much more to the canonical A-RNA values – this reflects both the improved quality of the diffraction data and the robustness of current refinement methods. Both of these new analyses find new magnesium-binding sites, and a greater number of well-defined water molecules hydrogen-bonding to the tRNA.

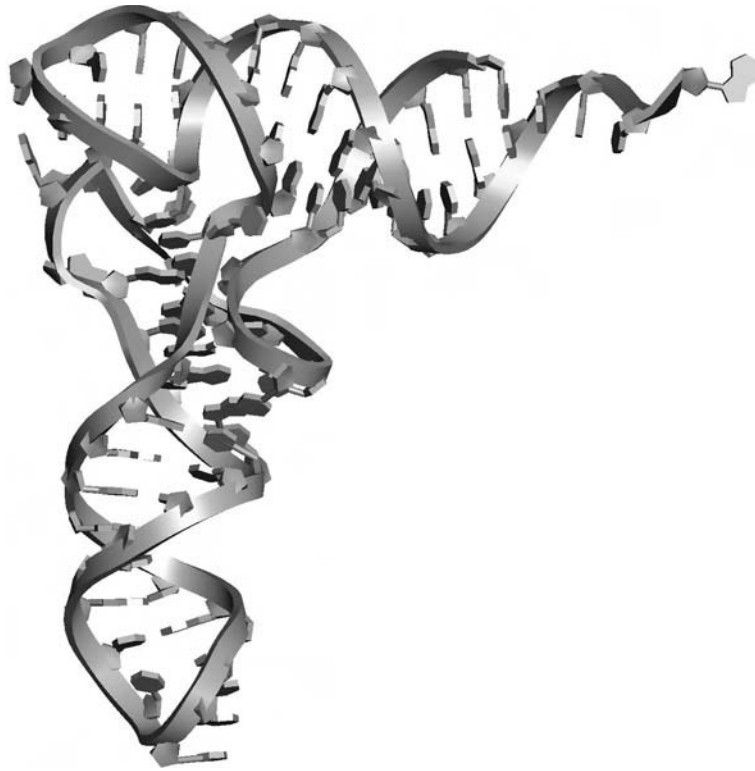


Figure 6.19 The structure of yeast tRNA^{Phe}, showing the L-shape (Robertus *et al.*, 1974; Kim *et al.*, 1974).

6.4 Ribozymes

The ability of certain RNA molecules to catalytically cleave either themselves or other RNAs is shown by a number of distinct categories of RNA molecule:

- The RNA of self-splicing group I introns—that from *Tetrahymena* was the first ribozyme to be discovered (Cech, Zang, and Gradowski, 1981; Guerrier-Takada *et al.*, 1983). These contain four conserved sequence elements and form a characteristic secondary structure. The initial step of the cleavage involves nucleophilic attack by a conserved guanosine.
- The RNA of self-splicing group II introns. These also have conserved sequence elements, but a very different secondary structure and a distinct mechanism of cleavage that involves the nucleophilic attack of a conserved adenosine, contained within the intron sequence.
- The RNA subunit of the enzyme RNase P
- Self-cleaving RNAs from viral and plant satellite RNAs. These are smaller ribozymes than the group I or II intron ones, and include the hammerhead ribozyme (Fig. 6.21)
- Ribosomal RNA in the ribosome, and in particular the peptidyl transferase machinery

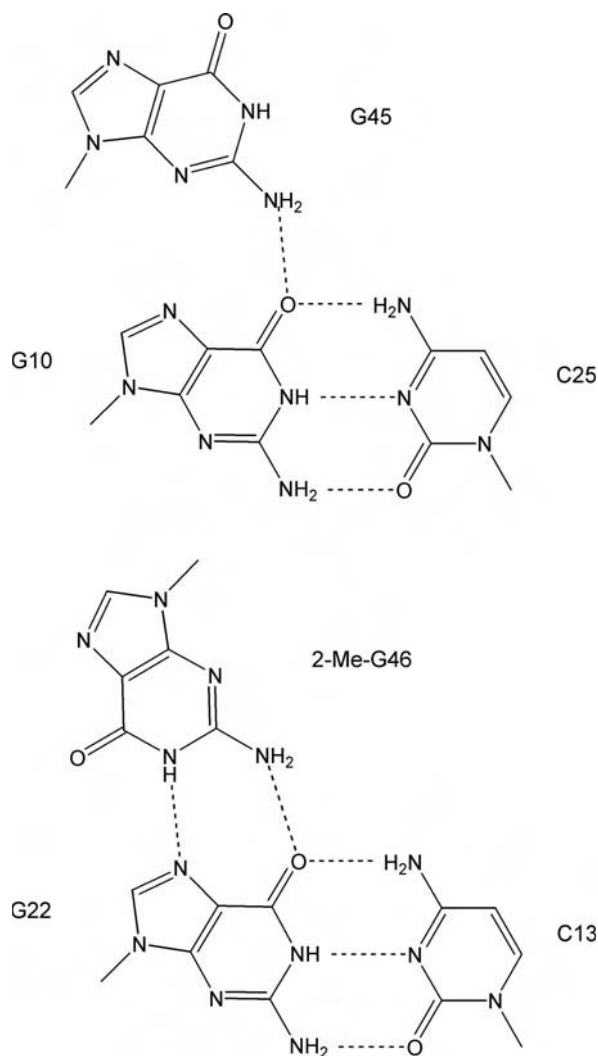


Figure 6.20 Base triplet interactions in the yeast tRNA^{phe} crystal structure.

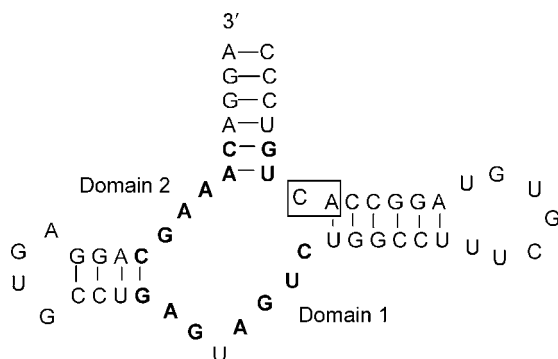


Figure 6.21 Sequence of a hammerhead ribozyme, indicating the hydrogen-bonded helical stems. The CA cleavage site is shown boxed.

6.4.1 The Hammerhead Ribozyme

The crystal structures of hammerhead ribozymes were the first to be determined, simultaneously by two groups (Pley, Flaherty, and McKay, 1994; Scott, Finch, and Klug, 1995). Both structures are similar in their essential detail. One (Pley, Flaherty, and McKay, 1994) has a DNA strand containing the putative cleavage site. Since ribozymes do not cleave DNA, this is effectively an inhibitor complex. The other (Scott, Finch, and Klug, 1995) is all-RNA, with a 2'-O-methyl group replacing the 2'-hydroxyl group of the active-site cytosine. The structures show three A-RNA stems connected to a central two-domain region containing the conserved residues (Fig. 6.22). The overall arrangement resembles a wish-bone. The two-residue sequence CA (the site of autocatalytic cleavage) is at the apex of two stems (Fig. 6.21), where there is a sharp turn in backbone conformation, that is focused on the uridine in domain one, formed by the invariant tetranucleotide sequence CUGA. This U-turn conformation is closely similar to those in the anticodon and pseudouridine loops of tRNA. The second domain, at the junction of two stems, contains conserved sheared tandem A•G base pairs. This performs the key structural role of maintaining the correct geometric relationships between cleaved and catalytic strands.

The overall mechanism of hammerhead ribozyme catalysis involves nucleophilic attack by a 2'-hydroxyl group on the phosphorus atom of the adjacent backbone (Fig. 6.23). However elucidation of the details of the mechanism has not been straightforward, although it was early on recognized that metal ions often play an important role, in view of the requirement for magnesium ions. The location of magnesium ions close to the A•G base pairs suggests that they provide a directly stabilizing influence on the precise conformation needed for catalysis. Several of the models for catalysis that have been proposed (Lilley, 1999), suggest that the ions are more directly involved.



Figure 6.22 The crystal structure of a hammerhead ribozyme (Pley, Flaherty, and McKay, 1994; Scott, Finch, and Klug, 1995).

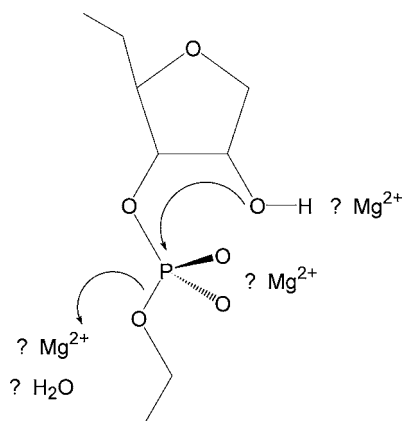


Figure 6.23 Schematic view of possible mechanisms of backbone cleavage by a hammerhead ribozyme. Sites of possible magnesium ion and water molecule involvement are indicated.

Magnesium ion(s) have been located at various positions within the active site, suggestive of a role in aiding the movement of electrons, perhaps by stabilization of cleaving groups. Several crystal structures, together with evidence from other biophysical methods, have provided data relevant to the mechanistic questions (Murray *et al.*, 1998; Feig, Scott, and Uhlenbeck, 1998; Murray *et al.*, 2000). Terbium ions inhibit catalytic cleavage, and the structure of a terbium complex (Feig, Scott, and Uhlenbeck, 1998) shows this ion to be located close to residues G5 and A6, and not within the active site. The structure of an intermediate has been captured (Murray *et al.*, 2000), which suggest that there is large-scale conformational change during the reaction, that brings the cleavage site and G5/A6 closer together than found in the ground-state structures (Pley, Flaherty, and McKay, 1994; Scott, Finch, and Klug, 1995).

6.4.2 Complex Ribozymes

The crystal structure of a Group I ribozyme domain, has been determined, that of the P4-P6 domain of the *Tetrahymena* intron (Cate *et al.*, 1996, and recently the complete intron from several different organisms (Adams *et al.*, 2004; Guo, Golden, and Cech, 2004; Golden, Kim, and Chase, 2005) shows an architecture formed by ten helices held together by the intervening loops and junctions, with the catalytic site close to a pseudoknot and two junction regions (Vicens and Cech, 2005).

The structure of the *Tetrahymena* intron (Fig. 6.24), contains 160 nucleotides, and is arranged as two roughly colinear helical regions. Within these are a number of RNA motifs, notably an adenosine-rich bulge, a three-way junction, and a GAAA tetraloop. The close packing of the helices in this and other large RNA structures is due to the highly favorable interactions between unpaired, flipped-out adenosines attached to one helix and the minor groove of the adjacent one (Doherty *et al.*, 2001). This provides an explanation for the high abundance of conserved unpaired adenosine residues in a number of large RNA molecules.

The crystal structures of the hepatitis delta virus and a hairpin ribozyme both show features relevant to their RNA-cleavage activities. The hepatitis delta virus (Ferré-D'Amaré, Zhou, and Doudna, 1998) has a very complex arrangement of five helical

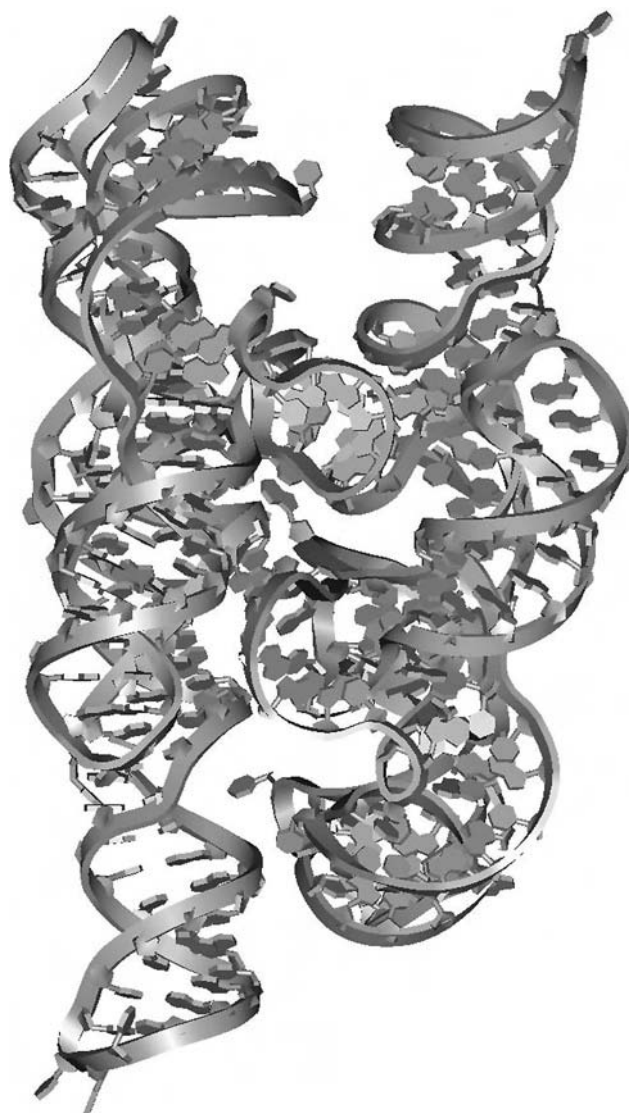


Figure 6.24 The structure of the P4-P6 domain of the *Tetrahymena* intron (Cate *et al.*, 1996).

moieties forming a double pseudoknot arrangement (Fig. 6.25). (The pseudoknot is another recurrent motif in RNA structure that involves the hydrogen-bonded interaction of a single-stranded region with a loop (Fig. 6.26)). This arrangement is fundamentally distinct from that in other ribozymes, as is the active site, which is buried within the five-helical core (Fig. 6.27). Another complex catalytic site is seen in the structure of the hairpin ribozyme (Rupert and Ferré-D'Amaré, 2001); both have features resembling enzymes that also use activated phosphates, such as ribonuclease S (Strobel and Ruder, 2001). The hairpin ribozyme has a four-way junction forming the central feature of the structure. The structures of several of the individual stems forming the junction have been previously individually studied; it is a salutary lesson in the complexities of RNA structure that these have undergone large conformational changes when a part of the complete hairpin ribozyme.

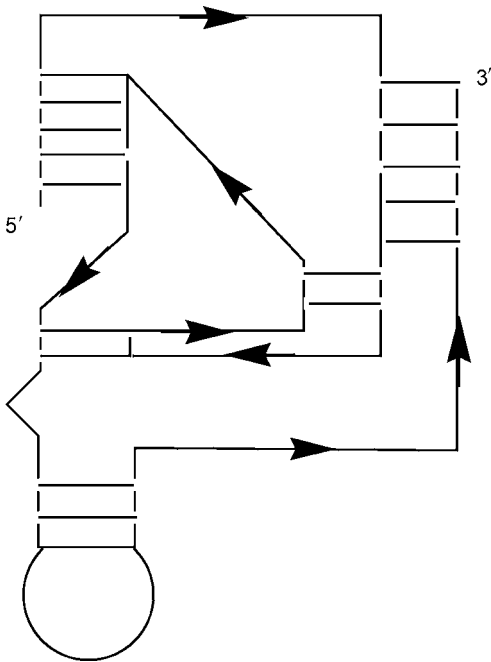


Figure 6.25 Topology of the hepatitis delta virus ribozyme (Ferre-D'Amare, Zhou, and Doudna, 1998).

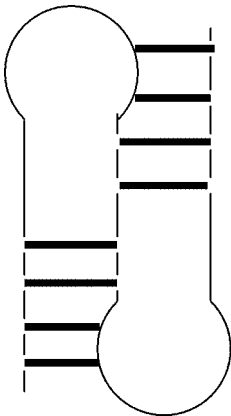


Figure 6.26 Topology of a pseudoknot.

Table 6.5 Ribozyme Crystal Structures		
Ribozyme type	PBD code no	NDB code no
RNA Hammerhead	299D	URX057
P4–P6 domain	1GID	URX053
<i>Tetrahymena</i> ribozyme	1GRZ	UR0003
Hepatitis delta virus	1DRZ	PR0005
Hairpin	1M5K	PR0038
RNA part of ribonuclease P	2A2E; 2A64	UR0064; UR0065
Group I intron	1ZZN	PR0170
Ligase ribozyme	2OIU	UR0110

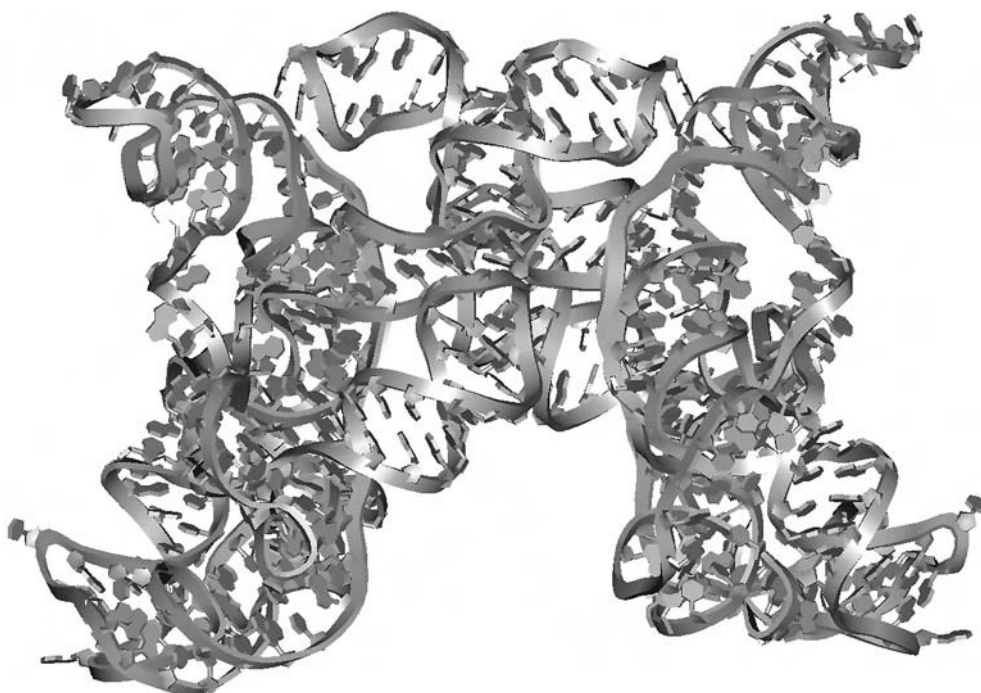


Figure 6.27 Crystal structure of the hepatitis delta virus ribozyme (Ferre-D'Amare, Zhou, and Doudna, 1998).

6.5 Riboswitches

The surprising finding that sequences in the 5' nontranslated region of certain bacterial mRNAs can act to regulate gene expression in the absence of proteins, is of major functional (and potentially therapeutic) significance. They do this by undergoing major conformation changes under the action of appropriate endogenous small-molecule metabolites (Winkler *et al.*, 2004). A very recent finding has been that riboswitches may also function in eukaryotic cells (Cheah *et al.*, 2007), where they have been shown to regulate RNA-splicing. Riboswitches can act by binding to, in particular, purine bases, some enzyme cofactors, and particular amino acids.

The majority of riboswitches have two key domains, an aptamer component for ligand-binding, and an “expression platform,” which transmits the ligand-induced conformational change to result in changes in gene expression. This is shown, for example, in the crystal structure of the guanine-responsive riboswitch from the *xpt-pbuX* operon of the *B. subtilis* organism, bound to the guanine-like base hypoxanthine (Batey, Gilbert, and Montange, 2004). This has the loops at the top end of two parallel helices P1 and P2 linked together by a number of base–base hydrogen bonds; at the other end the binding pocket is formed at the three-way junction with the third helix P3 (Figs. 6.28 and 6.29). The ligand is almost totally enclosed, and as shown in the guanine-bound riboswitch (Serganov *et al.*, 2004), is extensively hydrogen-bonded to the surrounding RNA residues, which are highly conserved (Fig. 6.30). This type of riboswitch turns off gene expression. However some riboswitches can activate gene expression, possibly by only allowing the transcription terminator stem

form in the absence of, for example, adenine (Serganov *et al.*, 2004). The guanine and adenine riboswitches share a common tertiary structure even though the extent of overall sequence identity is 59%.

A more complex category of riboswitch binds the coenzyme thiamine pyrophosphate, the so-called *thi*-box. Three independent crystal structures have been determined



Figure 6.28 The structure of the hypoxanthine riboswitch (Batey, Gilbert, and Montange, 2004).

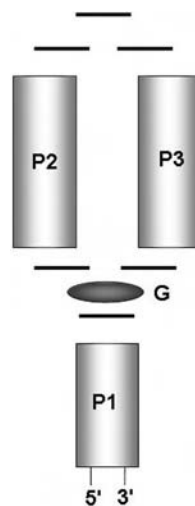


Figure 6.29 Schematic view of a simple riboswitch, with the three helices P1, P2, and P3 shown. Bound purine is also shown schematically as a shaded disc.

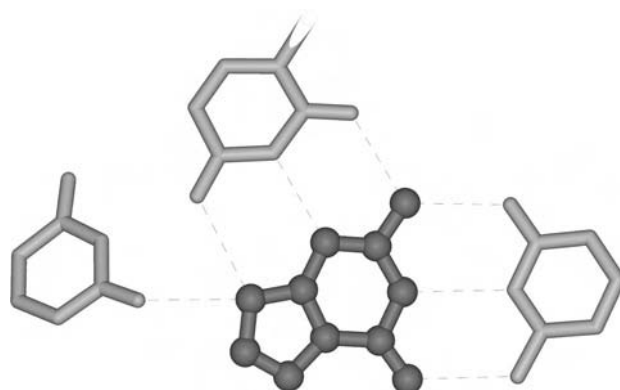


Figure 6.30 Detailed view of the bound guanine base, in the guanine riboswitch (Serganov *et al.*, 2004), with hydrogen bonds shown as dashed lines.

Table 6.6 Selected Riboswitch Crystal Structures

Riboswitch type and ligand	PBD code no	NDB code no
S-adenosylmethionine riboswitch	2GIS	UR0082
Guanine riboswitch + guanine	1Y27	UR0052
Guanine riboswitch + hypoxanthine	1U8D	UR0039
Thiamine pyrophosphate riboswitch + thiamine diphosphate	2CKY	UR0098
	2H0J	UR0099

for this riboswitch (Edwards and Ferré-D'Amaré, 2006; Serganov *et al.*, 2006; Thore, Leibundgut, and Ban, 2006). All are closely similar (apart from small sequence changes in nonconserved regions) and show that the Y-shaped molecule comprises two irregular RNA helices/loops linked by a three-way junction to a third helix (Fig. 6.31). The ligand-binding region is in a bulged pocket lined by conserved residues, and is held in place by a network of hydrogen bonds to these bases (Fig. 6.32), as well as interactions between some of the metal cations in the structure and the pyrophosphate tail.

6.6 The Ribosome, a Ribozyme Machine

The ribosome is responsible for protein synthesis in all prokaryotic and eukaryotic cells. It consists of two subunits, each a complex of proteins and ribosomal RNA. The complete 70S prokaryotic ribosome has a total molecular weight of ~2.5 million Daltons. In prokaryotes the 30S subunit contains about 20 proteins and a single RNA molecule of around 1500 nucleotides in length. The larger 50S subunit contains half as many more proteins and a large RNA of about 3000 nucleotides, together with the small (120 nucleotide) 5S RNA. The primary function of the small subunit is to control tRNA interactions with messenger RNAs. The large subunit controls peptide transfer and undertakes the catalytic function of peptide bond formation.

Structural studies on bacterial ribosomes have been underway for almost 50 years, with the ultimate goal of achieving atomic resolution in order to understand the mechanics of ribosome function – for most of that time appearing to be an impossibly difficult task. A number of individual ribosomal proteins were initially studied, but their three-dimensional arrangement within the ribosome, and their functional roles, were not apparent.



Figure 6.31 Structure of the thiamine pyrophosphate riboswitch, with the thiamine pyrophosphate shown in ball and stick form (Edwards and Ferré-D'Amaré, 2006).

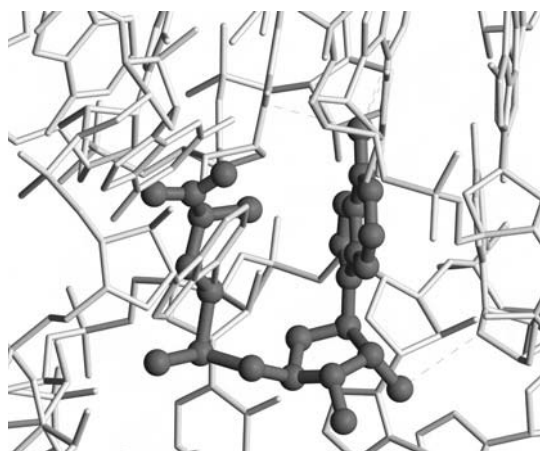


Figure 6.32 An enlarged view of the bound thiamine pyrophosphate in the riboswitch complex, with hydrogen bonds shown as dashed lines.

Electron microscopy has enabled the outlines of ribosome structure to be defined, which has been of considerable help to the more recent single-crystal analyses. Crystals of diffraction quality were first obtained well over 20 years ago, but only relatively recently have the various necessary technologies become available for tackling such a massive and complex crystallographic task. It was necessary to collect huge data sets from the very large unit cells involved, with crystals that diffracted only weakly and decayed rapidly in the X-ray beam. This has been achieved using synchrotron radiation together with low-temperature methods to stabilize the crystals from excessive decay.

Determining the phases of these large structures has also been a formidable challenge. Rather than using single heavy atoms, the use of clusters of heavy metals has been successful, especially for the initial structures, at less than atomic resolution. The progress of ribosome structure determination has been rapid (Table 6.7). The first structure, at 9 Å resolution, was reported in 1998 (Ban *et al.*, 1998), culminating in electron-density maps of the small (Tocija *et al.*, 1999) and large (Yusupov *et al.*, 2001) subunits at 3.0 Å and 2.4 Å resolution respectively, at which point individual side-chains and bases become apparent and can be refined. Recently, several complete ribosomal structures have been determined, of the *E. Coli* ribosome (Schuwirth *et al.*, 2005), and from *T. thermophilus* with tRNA (Korostelev *et al.*, 2006) and with tRNA and mRNA (Selmer *et al.*, 2006). These structures are some of the largest determined by X-ray crystallography, and dissection of the electron density maps has only been possible with the aid of the large body of biochemical and other data accumulated on the ribosome and its subunits. It is remarkable that the secondary structures for the RNAs in these complexes are close to those predicted by analyses of RNA across species, using the methods of phylogenetic analysis, which are based on the reasonable assumption that RNA structure is largely conserved throughout evolution.

A number of ribosome coordinate data sets are now available from the Data Banks. Although only backbone C_α carbon atoms were deposited for the protein components in the earlier lower-resolution structures, protein side-chain atoms have been identified in the more recent analyses, as well as complete coordinate sets for the RNA components. The sheer size of the PDB files involved has resulted in a number being split over more than one entry – this is reflected in Table 6.8. The size and complexity of ribosome structures is such that there is some advantage when viewing these very large assemblies, in restricting initial viewing to backbone atoms. The necessarily brief survey below of the structures cannot do justice to their complexity, and the reader is strongly encouraged to read the primary literature on them.

Table 6.7 Progress in Ribosome Crystallographic Structure Determination

Subunit	Organism	Resolution (Å)	Reference
50S	<i>H. marismortui</i>	9	Ban <i>et al.</i> , 1998
50S	<i>H. marismortui</i>	5.0	Ban <i>et al.</i> , 1999
30S	<i>T. thermophilus</i>	5.5	Clemons <i>et al.</i> , 1999
70S	<i>T. thermophilus</i>	7.8	Cate <i>et al.</i> , 1999
30S	<i>T. thermophilus</i>	4.5	Tocija <i>et al.</i> , 1999
30S	<i>T. thermophilus</i>	3.0	Wimberley <i>et al.</i> , 2000
30S	<i>T. thermophilus</i>	3.3	Ban <i>et al.</i> , 2000
50S	<i>H. marismortui</i>	2.4	Yusupov <i>et al.</i> , 2001
70S	<i>T. thermophilus</i>	5.5	Carter <i>et al.</i> , 2000
70S	<i>T. thermophilus</i>	3.7	Korostelev <i>et al.</i> , 2006
70S	<i>T. thermophilus</i>	2.8	Selmer <i>et al.</i> , 2006
70S	<i>E. Coli</i>	3.5	Schuwirth <i>et al.</i> , 2005

Table 6.8 Selected Ribosome Subunit Crystal Structures

Subunit and resolution	PDB code no	NDB code no
30S, 3.0 Å	1FJG	RR0016
50S, 2.4 Å	1GIX	RR0031
30S with mRNA + tRNA, 3.3 Å	1IBL, 1IBM	RR0029, RR0030
70S+ tRNA + mRNA, 5.5 Å	1GIX, 1GIY	RR0031, RR0032
70S + tRNA, 3.71 Å	2OW8, 1VSA	RR0163, RR0164
70S + tRNA + mRNA, 2.8 Å	2J01, 2J02, 2J03, 2J04	RR0154-RR0157
70S ribosome from <i>E. Coli</i>	2AVY, 2AW4, 2AW7, 2AW8	RR0123-RR0126

The initial crystal structure of the complete 70S ribosome (Yusupov *et al.*, 2001), although at relatively low resolution, revealed much about the interactions between the subunits that are an essential element of the protein-synthesis cycle. It is notable that even though the resolution precluded detailed study of the interactions involved, the bound tRNA molecules were all seen to be extensively contacted by ribosomal RNA in addition to the necessary established functional interactions such as codon-anticodon recognition. More recent analyses, at 2.8 Å (Selmer *et al.*, 2006) and at 3.7 Å (Korostelev *et al.*, 2006), together with that of the complete *E. Coli* ribosome at 3.5 Å (Schuwirth *et al.*, 2005), have enabled atomic-level details to be unambiguously defined.

6.6.1 The Structure of the 30S Subunit

The 3.0 Å crystal structure (Wimberley *et al.*, 2000), together with the more recent higher-resolution analyses, have located the complete 16S ribosomal RNA of 1511 nucleotides together with the ordered regions of 20 ribosomal proteins, altogether organized into four well-defined domains. The implication is that there is considerable flexibility between them, which is needed in order to ensure the movement of messenger and transfer RNAs. The overall shape of the 30S particle is dominated by the structure of the folded RNA (Fig. 6.33), and the proteins serve merely to fill up the gaps and hold the whole assembly together. The secondary structure of the RNA shows a very large number (>50) helical regions. The numerous loops are mostly small and do not disrupt the runs of helix in which they are embedded. There are extensive interactions between helices, mostly involving coaxial-stacking via the minor grooves. In one type of helix-helix interaction two minor grooves abut each other, with consequent distortions from A-type geometry. These distortions, which tend to involve runs of adenines, are facilitated by both extrahelical bulges and noncanonical base pairs, as have been observed in simple RNA structures. Less commonly, perpendicular packing of one helix against another (also via the minor groove) is mediated by an unpaired purine base. This mode is of special importance since it involves the functionally significant helices in the 30S subunit. The motifs of RNA tertiary structure such as non-Watson-Crick base pairs, base triplets and tetraloops, all contribute to the overall structure.

The majority of the 20 ribosomal proteins in the structure each consist of a globular region and a long flexible arm. The latter have been too flexible to be observed in structural studies on the individual proteins, but have been located in the 30S subunit, where they play important roles in helping to stabilize the RNA folding, by essentially filling in the numerous spaces in the RNA folds.

The structure identifies the three sites where tRNA molecules bind and function, and where the essential proofreading checks for fidelity of code-reading and translation occur. These sites are:

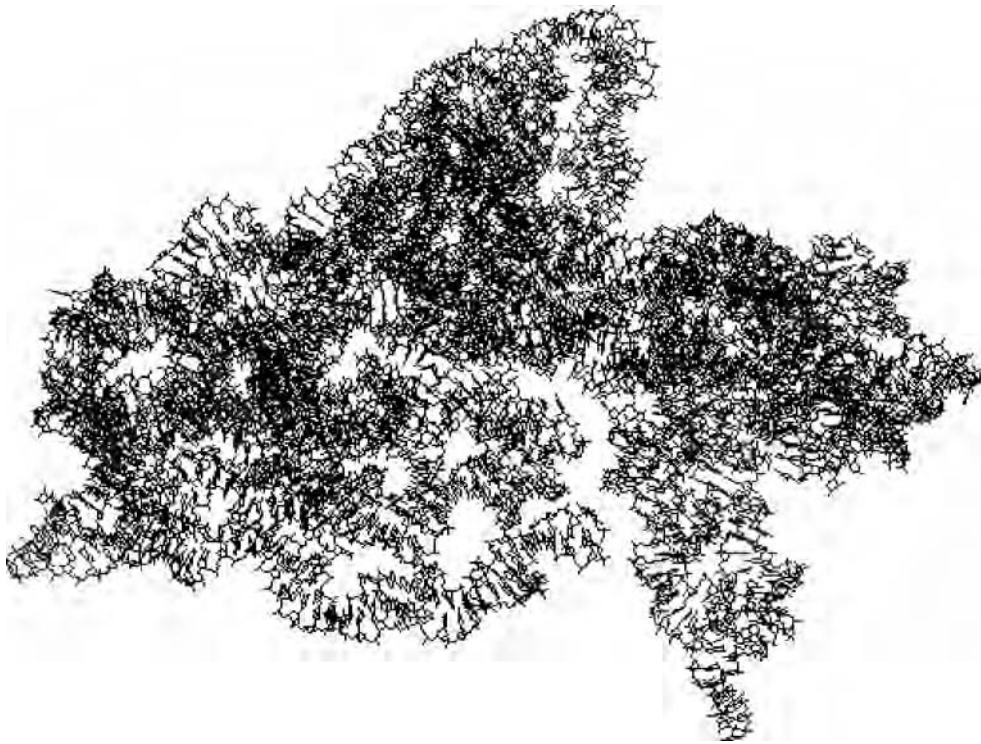


Figure 6.33 Crystal structure of the RNA component in the 30S ribosome subunit (Wimberley *et al.*, 2000).

- The P (peptidylation) site, when a tRNA anticodon base pairs with the appropriate codon in mRNA, and where the peptide chain is covalently linked to a tRNA
- The A (acceptor) site, when peptide bonds are eventually formed (the actual peptidyl transferase steps occur in the 50S subunit)
- The E (exit) site, for tRNAs to be released from the subunit as part of the protein-synthesis cycle.

tRNA itself was not present in the first 30S crystal structure, but the RNA from a symmetry-related 30S subunit effectively served to mimic it as the anticodon stem-loop. Its interactions with the 30S RNA are also mediated via (i) minor-groove surfaces, helped by some contacts with ribosomal proteins, and (ii) via backbone contacts. Interestingly, the exit site of the tRNA is almost exclusively protein-associated, whereas the other functional sites are composed of RNA and not protein. It is thus the ribosomal RNA that mediates the functions of the 30S subunit, and not the ribosomal proteins.

The determination (Ogle *et al.*, 2001) of the crystal structure at 3.3 Å of the 30S subunit complete with the anticodon stem-loop of a tRNA and a short RNA acting as a mRNA, has enabled the details of codon-anticodon recognition and checking to be visualized. A key role is played by two adjacent adenines from the 16S RNA, which are in contact with the minor groove of the codon-anticodon helical stem, so as to verify the correctness of the first two Watson–Crick base pairs of the triplet, and selection of the correct codon. The third codon position (the “wobble” position), is not as involved in interactions with the 16S RNA and a tRNA anticodon is thus able to cope with hydrogen-bonding to a range of 3rd-position codon bases. The codon-anticodon

stringency is disrupted by the antibiotic paromomycin, which has also been cocrystallized with the 30S subunit.

Several other antibiotics that are known to inhibit protein synthesis have been examined bound to the 30S subunit, following cocrystallization experiments (Carter *et al.*, 2000; Brodersen *et al.*, 2000). For example, streptomycin has been found to bind tightly to four distinct sites on the 16S RNA, mostly via phosphate groups. This stabilizes the A site in an open conformation, thereby enabling noncognate tRNAs to bind and reducing proofreading capability. These and other studies are relevant to the structure-based design of new antibiotics to be rationally designed, and are further discussed in Sect. 6.5.2. A second crystal structure (Schlunzen *et al.*, 2000), also at 3.3 Å resolution, has shown that the decoding region of the structure, is highly conserved. Furthermore the A and P site tRNAs and the codon part of the bound mRNA do not interact with any of the proteins, again emphasizing the dominant role of RNA in ribosome structure and function.

6.6.2 The Structure of the 50S Subunit

In the 2.4 Å crystal structure (Ban *et al.*, 2000) 2711 out of the total of 2923 ribosomal RNA nucleotides have been observed, together with 27 ribosomal proteins, and 122 nucleotides of 5S RNA, with the subunit being about 250 Å in each dimension. Again, the function of the proteins appears to be mainly to act as a cement, helping to maintain the integrity of the folded RNA. Although the RNA itself can be arranged into six domains on the basis of its secondary structure, overall the 50S subunit is remarkably globular, reflecting its greater conformational rigidity compared to the 30S subunit, consist with its functional need to be more flexible.

The principal function of the 50S subunit is peptide bond formation. The active site where this occurs is within domain 5. Its precise location was identified by the analysis of crystals containing the antibiotic puromycin, an inhibitor of peptide synthesis. Remarkably, there is no protein structure within 18 Å of this molecule, bound in its active site, and accordingly it has been proposed that peptidyl transferase catalytic activity is entirely carried out by RNA, analogous to the function of RNAs as ribozymes. The reaction is an acid-base hydrogen transfer, with an adenine being suggested as the key residue, accepting a hydrogen atom from an aminoacylated tRNA so that it can bond to the carbonyl group of an esterified peptidyl-tRNA carrying the peptide chain to which the new amino acid is to be attached. All nucleotides involved in the active site are highly conserved in nature, strongly suggesting that RNA catalysis is a fundamental mechanism acting throughout evolution.

6.6.3 Complete Ribosome Structures

Structural studies on complete ribosomes are important for full understanding of mechanism since they do not function as individual subunits, but as complete 70S particles. The crystals of the *E. Coli* ribosome (Schuwirth *et al.*, 2005) have provided two structures for this intact ribosome, since there are two independent ribosome assemblies per crystallographic asymmetric unit. There are some significant differences between the two structures, notably that the 30S subunit is rotated about 6° around the so-called neck region (which connects the 30S and 50S subunits), in one structure

relative to the other, and is also different from the conformation seen in the 2.8 Å crystal structure of the intact 70S *T. thermophilus* ribosome (Selmer *et al.*, 2006), in which tRNA molecules are clearly seen (Fig. 6.34) together with part of the mRNA (some of it is disordered and is not visible in the electron density). Figures 6.35 and 6.36 shows the high level of detail observed in this structure.

6.7 RNA-Drug Complexes

A variety of RNA molecules and assemblies are targets for therapeutic intervention by small molecules. An important target that has received much attention, is the HIV Trans-activator of transcription (Tat) protein and Trans Activation Responsive region (TAR) RNA. This has an arginine-rich group Tat-recognizing the base sequence and the conformation of TAR RNA. NMR studies have been extensively used for this system, and has, for example, identified the binding site of the drug acetylpromazine on TAR RNA (Du, Lind, and James, 2002).



Figure 6.34 (a–c) Three views of the 30S ribosomal subunit, as found in the crystal structure of the complete 70S particle (Selmer *et al.*, 2006). The protein secondary structure elements are shown colored dark grey. The RNA backbones are shown in tube form and bases are represented as sticks.

(Continued)

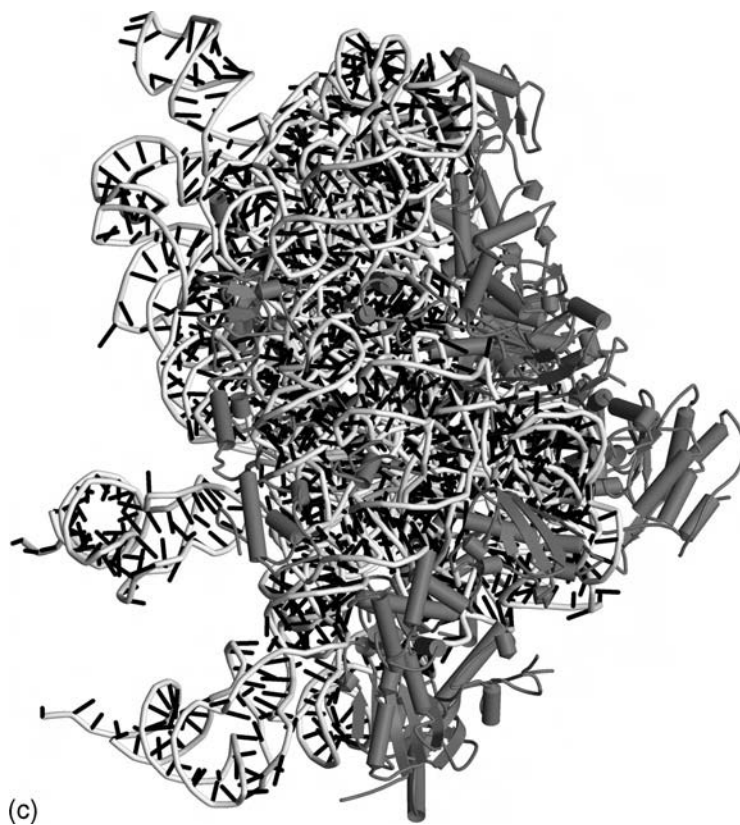
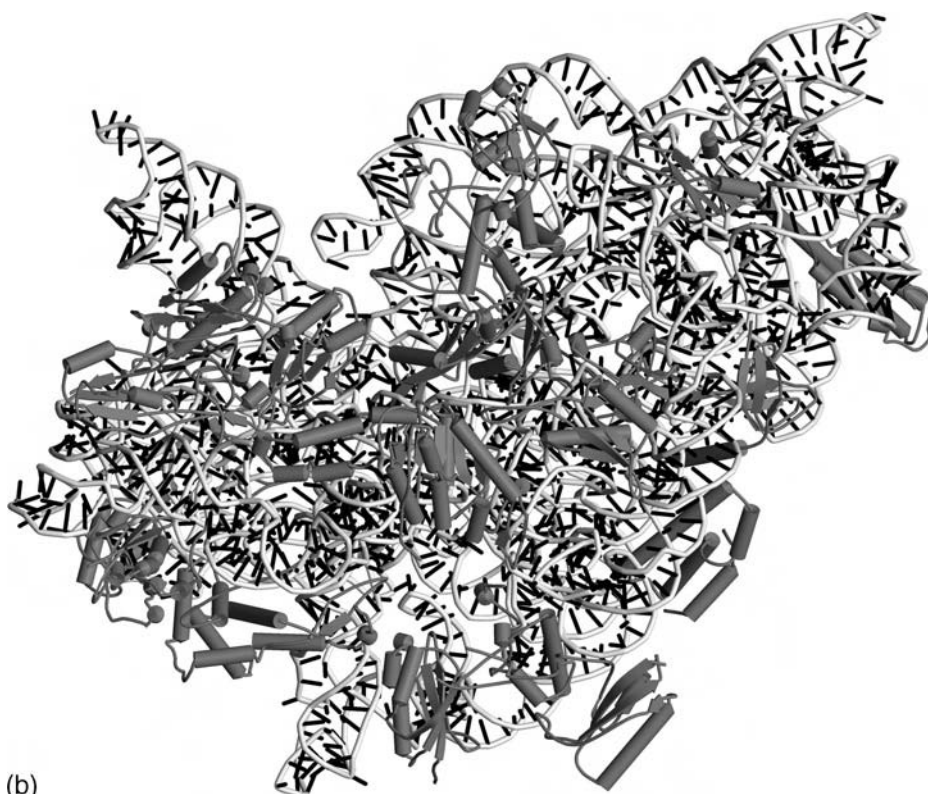


Figure 6.34 Cont'd.



Figure 6.35 The RNA components of the 30S ribosomal subunit, as found in the crystal structure of the complete 70S particle (Selmer *et al.*, 2006). Only the backbone is shown for the ribosomal and mRNAs; the incoming and outgoing tRNA molecules are seen at bottom left.

Many antibacterial antibiotics act directly at the ribosomal level, binding to RNA structural features and thereby blocking correct translation (Böttger, 2006; Poehls-gaard and Douthwaite, 2005). As outlined in the previous section, there have been a number of crystal structure analyses of ribosome structures with bound drugs. For example, the crystal structures of complexes of the large ribosomal subunits from the organism *H. marismortui* (Tu *et al.*, 2005) with bound erythromycin, azithromycin, clindamycin, virginiamycin S, and telithromycin, have provided some insights into the resistance mechanisms involving these antibiotics. These all contain the mutation G2099A, which confers increased affinity on erythromycin, possibly as a result of the loss of solvent at the N2 position on going from G to A. Crystallographic analyses (Hansen *et al.*, 2002) of the four macrolide antibiotics (carbomycin A, spiramycin, tylosin, and azithromycin) with the same subunit have shown that these drugs bind in the polypeptide tunnel adjacent to the peptidyl transferase center, and therefore that their action involves blocking a growing peptide chain from exiting.

Since the crystal structure of the ribosomal subunit for a particular microorganism is not always available, or is often at only moderate resolution (ca 3 Å), a more rapid and

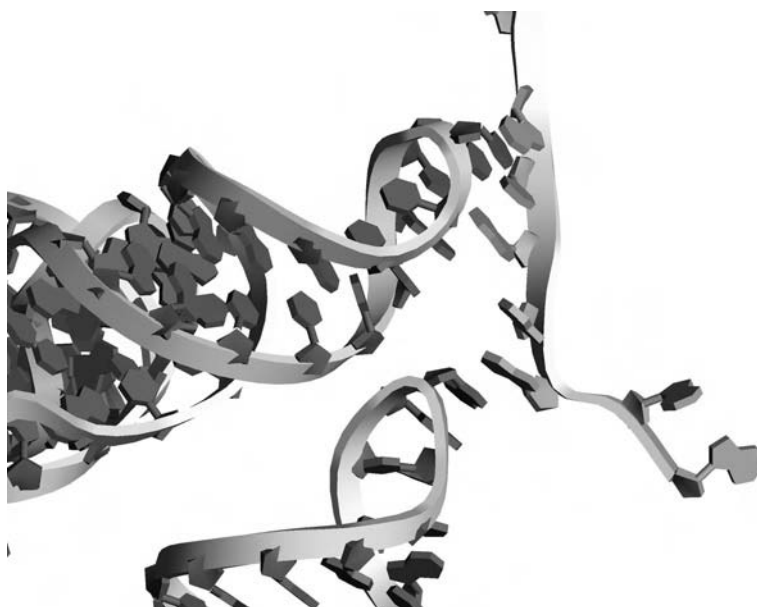


Figure 6.36 Detail of the interactions (Selmer *et al.*, 2006) between the anticodon stems of the two tRNA molecules and codons in a fragment of mRNA (on the right-hand side, adopting an extended conformation).

precise approach to antibiotic drug design has been to analyze cocrystals between the antibiotic and the RNA that constitutes its minimal binding site. The power of this approach has been shown by the analyses of complexes involving the aminoglycoside antibiotics gentamicin C1A, kanamycin A, ribostamycin, lividomycin A, neomycin B (François *et al.*, 2004), and several paromomycin derivatives (Hanessian *et al.*, 2007). This latter study itself follows on from earlier analyses (Vicens and Westhof, 2001) of a complex of the native paromomycin antibiotic and two synthetic derivatives (François *et al.*, 2004) with the decoding A-site RNA of *E. Coli*. Paromomycin itself is in an L-shape when bound in this site, with a large number of hydrogen-bond and electrostatic contacts (Fig. 6.37), that are shared with most of the A-site bound aminoglycoside antibiotics. It was reasoned (François *et al.*, 2004) that appropriate chemical modification of paromomycin would lead to a distinct bound conformation and set of contacts, with enhanced potency and activity against resistant strains, as was subsequently found experimentally and structurally. Structure-based methods have also been successfully employed (Fig. 6.38) in the design of three synthetic derivatives of the antibiotic enamine, which similarly binds to the A-site (Murray *et al.*, 2006). Confirmation of the bound conformation was obtained by a crystal structure analysis of one derivative bound to a complete 30S ribosome subunit.

Table 6.9 Selected RNA-drug Crystal Structures

RNA type and drug	PDB code no	NDB code no
Neomycin + A-site	2ET4	DR0016
Paromomycin + A-site	1J7T	DR0006
Streptomycin + 40-mer aptamer	1NTB	DR0010
Designer antibiotic1 + A-site	2F4S	DR0027
Designer antibiotic2 + A-site	2F4T	DR0028
Designer antibiotic3 + A-site	2F4U	DR0029

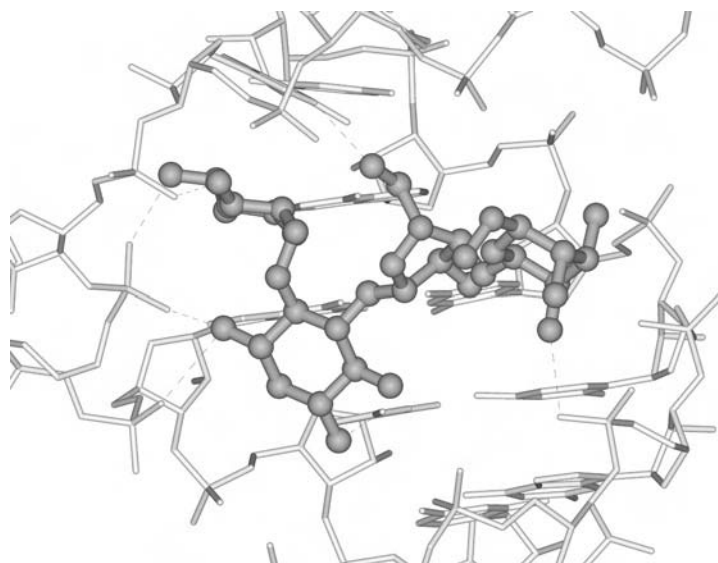


Figure 6.37 Paromomycin bound in the decoding A-site RNA of *E. Coli* this site, showing the hydrogen bond and electrostatic contacts (François *et al.*, 2004).

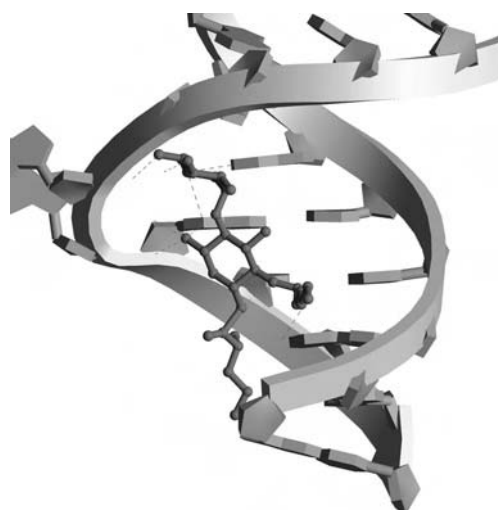


Figure 6.38 Crystal structure of a synthetic derivative of the antibiotic enamine, bound to the A-site (Murray *et al.*, 2006).

Streptomycin (Fig. 6.39) was the first antibiotic with activity against tuberculosis, although it is no longer the first-line treatment as a consequence of its side effects. It acts by binding to 16S ribosomal RNA, and a crystal structure of a complex with the 30S subunit has been determined (Carter *et al.*, 2000), which involves the active recognition of all three rings of the drug. Its cocrystal structure with a 40-mer RNA aptamer (Tereshko, Skripkin, and Patel, 2003), shows a very different arrangement, with an L-shaped RNA fold (Fig. 6.40). Only one part of the drug molecule (the six-membered carbohydrate streptose ring) is fully buried in the binding pocket that is formed from the two asymmetric internal loops of the RNA, and held by several hydrogen bonds.

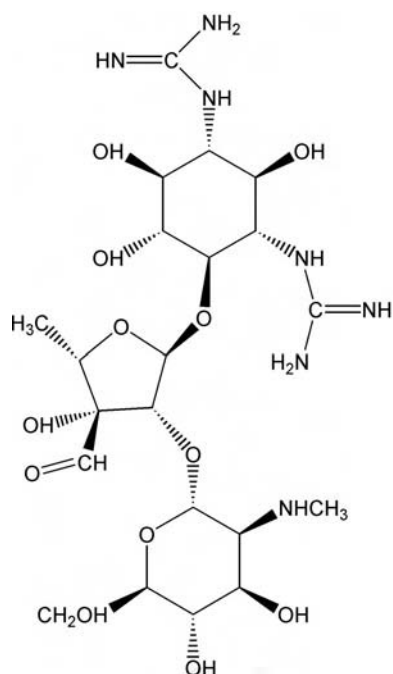


Figure 6.39 The structure of streptomycin.



Figure 6.40 (a,b) Two views of the streptomycin complex with a 40-mer RNA aptamer (Tereshko, Skripkin, and Patel, 2003). The enlarged view shows the hydrogen-bonding between the partially bound streptomycin molecule and RNA sites.

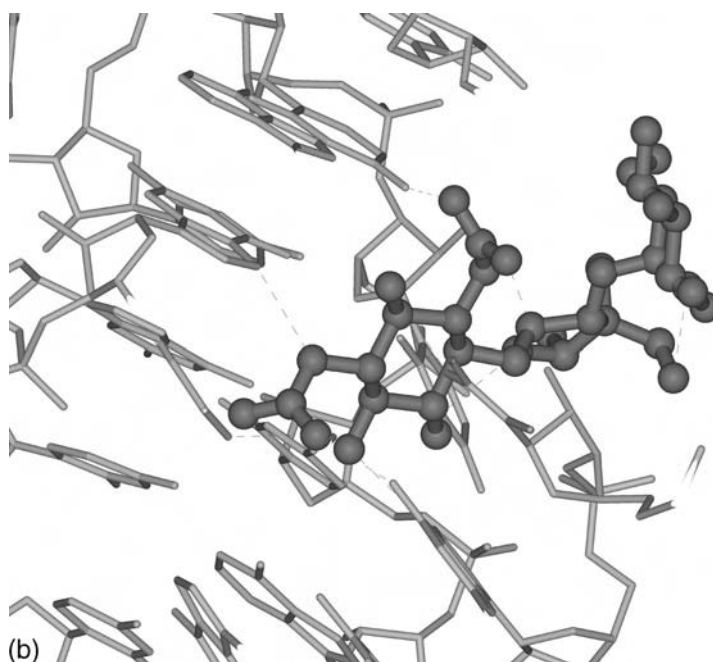


Figure 6.40 Cont'd.

6.8 RNA Motifs

We have seen in this chapter that the variability and complexity of RNA structures is due to their ability to fully exploit the hydrogen-bonding and base-stacking potential of both the bases themselves and the ribose sugars. In hydrogen-bonding terms Watson–Crick base-pairing remains the bedrock of helical stems, essential components of all RNAs. In addition, RNAs exploit the full gamut of potential base pairing presented by purines and pyrimidines, in ways that very much extend the mismatches seen with DNA, as discussed in Chap. 4. These pairings, which play key roles in stabilizing loop–loop interactions, have been systematized and classified into discrete families in a rational way (Leontis and Westhof, 2001; Leontis, Stombaugh, and Westhof, 2002; Lee and Gutell, 2004) that emphasizes the three edges of a base – Watson–Crick, Hoogsteen, and sugar/base. There is even greater diversity of possible base triplets and quartets, many of which have been observed in ribosome structures, which are the richest mines of information on RNA tertiary folds, as they have progressed to increasingly higher resolution. These interactions in turn are the basis of the various established RNA motifs, the adenosine platform, various tetraloops, the pseudoknot, the kink-turn and the C-loop (Lescoute *et al.*, 2005). It is already apparent that the sheer density of nucleotides in the ribosome results in a very large number of loop insertions into helices, primarily via their minor grooves, and a number of novel structural motifs are involved. One, the A-minor motif, appears to be especially prevalent (Nissen *et al.*, 2001). This involves adenines interacting in the minor groove, of which four distinct ways of doing so have been identified. The A-minor motif is an

important component of the manner in which ribosomal RNA preserves its functionally significant architecture – the adenines involved are often highly conserved, and so the ribosomal RNA framework is likely to be largely invariant throughout the living world. The motif also appears in several ribozyme structures (see also Sect. 6.4.2).

We have also seen that other fold motifs such as three-way junctions, play key roles in the architecture of ribozymes and riboswitches. The various structural types three-way junction can be categorized into three distinct topological families that depend on the size of the junction, and on the relative orientations of the three stems forming the arms of the junction (Lescoute and Westhof, 2006).

RNA motifs can also be categorized, for example, on the basis of conformational type (Schneider, Morávek, and Berman, 2004). An extension of this has been to use not single nucleotides, but libraries of pairs of structurally close nucleotides (Sykes and Levitt, 2005). From this, just 30 doublets have been found to represent most RNA structural (i.e. conformational and base-pairing) features.

The complexity of RNA structures inevitably means that the motifs for recognition of RNA structures in general by proteins are much more diverse than for DNA-protein recognition, which is dominated by the regularity of the DNA double helix. Where pure RNA helices are involved, proteins can insert α -helices, β -sheets or turns into the wide minor groove of an A-helix (Draper, 1999), and a number of preferred folds have been identified. Some of this variety in RNA-protein interaction modes

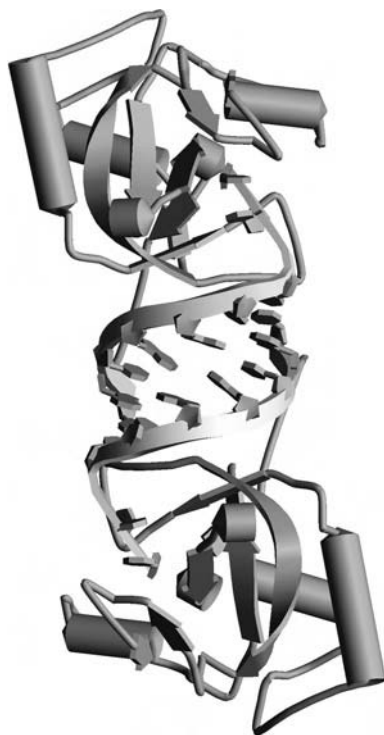


Figure 6.41 The crystal structure of the dimer formed by the PAZ domain, with a short, mostly duplex RNA at the interface between the two PAZ molecules (Ma, Ye, and Patel, 2004).

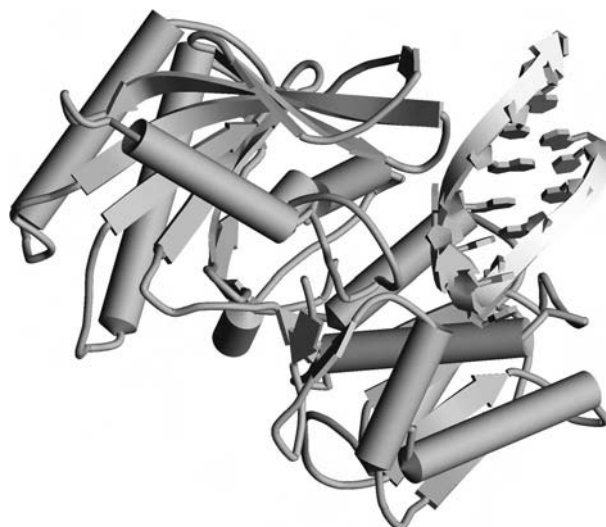


Figure 6.42 The crystal structure of the PIWI complex with a short RNA molecule bound (Parker, Roe, and Barford, 2005).

is illustrated by the crystal structure of a small RNA with a component of the siRNA machinery. In the 9-mer RNA complex with the PAZ domain of the Argonaute protein (Ma, Ye, and Patel, 2004), the RNA duplex in this dimer structure interacts with the C-terminus of the protein and a surface formed by β -sheet strands and some of the loops (Fig. 6.41). By contrast the short RNA in the PIWI complex (Parker, Roe, and Barford, 2005; Ma *et al.*, 2005), sits in a shallow L-shaped channel (Fig. 6.42) formed at the interface between the two domains of the protein.

6.9 Web Sites of Interest

<http://roc.bgsu.edu/> The RNA Ontology Consortium (Leontis *et al.*, 2006) provides further information on many aspects of RNA. This consortium is aiming to achieve definition and integration for all RNA structural and functional properties, not least RNA motifs and general structural features, such as have been discussed in this chapter.

<http://www-lbit.iro.umontreal.ca/mcsym/> This site gives access to the Mc-Sym and Mc-Annotate programs for building and analyzing RNA 3D structures (see Gendron, P., Lemieux, S. and Major, F. (2001). *J. Mol. Biol.*, 308, 919).

<http://www.rnabase.org> This site provides data on all publicly available three-dimensional RNA structures, taking coordinates that are deposited in the NDB. RNAbase also provides conformational maps and lists of parameter outliers.

http://prion.bchs.uh.edu/bp_type/ This site provides detail on noncanonical base pairs (Nagaswamy *et al.*, (2002). NCIR: a database of noncanonical interactions in known RNA structures. *Nucleic Acids Res.*, **30**, 395–397).

<http://riboswitch.bioapps.biozentrum.uni-wuerzburg.de/server.html>. Searches RNA (and DNA) sequences for potential riboswitches

References

- Adams, P. L., Stahley, M. R., Kosek, A. B., Wang, J., and Strobel, S. A. (2004). Crystal structure of a self-splicing group I intron with both exons. *Nature*, **430**, 45–50.
- Allain, F. H.-T. and Varani, G. (1995). Structure of the P1 helix from group I self-splicing introns. *J. Mol. Biol.*, **250**, 333–353.
- Anderson, A. C., Scaringe, S. A., Earp, B. E., and Frederick, C. A. (1996). HPLC purification of RNA for crystallography and NMR. *RNA*, **2**, 110–117.
- Arnott, S., Hukins, D. W. L., Dover, S. D., Fuller, W., and Hodgson, A. R. (1973). Structures of synthetic polynucleotides in the A-RNA and A'-RNA conformations: x-ray diffraction analyses of the molecular conformations of polyadenylic acid–polyuridylic acid and polyinosinic acid–polycytidylic acid. *J. Mol. Biol.*, **81**, 107–122.
- Arnott, S. (1999) In *Oxford Handbook of Nucleic Acid Structure* (Neidle, S., Ed.), p. **1**, Oxford University Press, Oxford.
- Auffinger, P. and Westhof, E. (1997). Rules governing the orientation of the 2'-hydroxyl group in RNA. *J. Mol. Biol.*, **274**, 54–63.
- Auffinger, P. and Westhof, E. (2000). Water and ion binding around RNA and DNA (C,G) oligomers. *J. Mol. Biol.*, **300**, 1113–1131.
- Baeyens, K. J., de Bondt, H. L., Pardi, A., and Holbrook, S. R. (1996). A curved RNA helix incorporating an internal loop with G.A and A.A non-Watson–Crick base pairing. *Proc. Natl. Acad. Sci. USA*, **93**, 12851–12855.
- Ban, N., Freeborn, B., Nissen, P., Penzyk, P., Grassucci, R. A., Sweet, R. M., Frank, J., Moore, P. B., and Steitz, T. A. (1998). A 9 Å resolution X-ray crystallographic map of the large ribosomal subunit. *Cell*, **93**, 1105–1115.
- Ban, N., Nissen, P., Hansen, J., Capel, M., Moore, P. B., and Steitz, T. A. (1999). Placement of protein and RNA structures into a 5 Å-resolution map of the 50S ribosomal subunit. *Nature*, **400**, 841–847.
- Ban, N., Nissen, P., Hansen, J., Moore, P. B., and Steitz, T. A. (2000). The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*, **289**, 905–920.
- Batey, R. T., Gilbert, S. D., and Montange R. K. (2004). Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine. *Nature*, **432**, 411–415.
- Böttger, E. C. (2006). The ribosome as a drug target. *Trends Biotechnol.*, **24**, 145–147.
- Brodersen, D. E., Clemons, W. M., Carter, A. P., Morgan-Warren, R. J., Wimberley, B. T., and Ramakrishnan, V. (2000). The structural basis for the action of the antibiotics tetracycline, pactamycin, and hygromycin B on the 30S ribosomal subunit. *Cell*, **103**, 1143–1154.
- Carter, A. P., Clemons, W. M., Brodersen, D. E., Morgan-Warren, R. J., Wimberly, B. T., and Ramakrishnan, V. (2000). Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. *Nature*, **407**, 340–348.
- Cate, J. H. and Doudna, J. A. (2000). Solving large RNA structures by X-ray crystallography. *Methods in Enzymology*, **317**, 169–180.
- Cate, J. H., Gooding, A. R., Podell, E., Zhou, K., Golden, B. L., Kundrot, C. E., Cech, T. R., and Doudna, J. A. (1996). Crystal structure of a group I ribozyme domain: principles of RNA packing. *Science*, **273**, 1678–1685.
- Cate, J. H., Yusupov, M. M., Yusupova, G. Z., Earnest, T. N., and Noller, H. F. (1999). X-ray crystal structures of 70S ribosome functional complexes. *Science*, **285**, 2095–2104.
- Cech, T. R., Zaug, A. J., and Grabowski, P. J. (1981). In vitro splicing of the ribosomal RNA precursor of Tetrahymena: involvement of a guanosine nucleotide in the excision of the intervening sequence. *Cell*, **27**, 487–496.
- Cheah, M. T., Wachter, A., Sudarsan, N., and Breaker, R. R. (2007). Control of alternative RNA splicing and gene expression by eukaryotic riboswitches. *Nature*, **447**, 497–500.
- Cheatham, T. E. and Kollman, P. A. (1997). Molecular dynamics simulations highlight the structural differences among DNA:DNA, RNA:RNA, and DNA:RNA hybrid duplexes. *J. Amer. Chem. Soc.*, **119**, 4805–4825.
- Clemons, W. M., May, J. L. C., Wimberley, B. T., McCutcheon, J. P., Capel, M. S., and Ramakrishnan, V. (1999). Structure of a bacterial 30S ribosomal subunit at 5.5 Å resolution. *Nature*, **400**, 833–840.
- Cruse, W. B. T., Saludjian, P., Biala, E., Strazewski, P., Prangé, T., and Kennard, O. (1994). Structure of a mispaired RNA double helix at 1.6-Å resolution and implications for the prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. USA*, **91**, 4160–4164.
- Dock-Bregeon, A. C., Chevrier, B., Podjarny, A., Johnson, J., de Bear, J. S., Gough, G. R., Gilham, P. T., and Moras, D. (1989). Crystallographic structure of an RNA helix: [U(UA)6A]₂. *J. Mol. Biol.*, **209**, 459–474.
- Dock-Bregeon, A. C., Moras, D., and Giegé, R. (1999). In *Crystallisation of Nucleic Acids and Proteins* (Ducruix, A. and Giegé, R., eds.), Oxford University Press, Oxford.
- Doherty, E. A., Batey, R. T., Masquida, B., and Doudna, J. A. (2001). A universal mode of helix packing in RNA. *Nature Struct. Biol.*, **8**, 339–343.
- Doudna, J. A., Grosshans, C., Gooding, A., and Kundrot, C. E. (1993). Crystallization of ribozymes and small RNA motifs by a sparse matrix approach. *Proc. Natl. Acad. Sci. USA*, **90**, 7829–7833.

- Draper, D. E. (1999). Themes in RNA-protein recognition. *J. Mol. Biol.*, **293**, 255–270.
- Du, Z., Lind, K. E., and James, T. L. (2002). Structure of TAR RNA complexed with a Tat-TAR interaction nanomolar inhibitor that was identified by computational screening. *Chem. Biol.*, **9**, 707–712.
- Edwards, T. E. and Ferré-D'Amaré. (2006). Crystal structures of the *thi*-box riboswitch bound to thiamine pyrophosphate analogs reveal adaptive RNA-small molecule recognition. *Structure*, **14**, 1459–1468.
- Egli, M., Minasov, G., Su, L., and Rich, A. (2002). Metal ions and flexibility in a viral RNA pseudoknot at atomic resolution. *Proc. Natl. Acad. Sci. USA*, **99**, 4302–4307.
- Egli, M., Portmann, S. and Usman, N. (1996). RNA hydration: a detailed look. *Biochemistry*, **35**, 8489–8494.
- Ennifar, E., Walter, P., Ehresmann, B., Ehresmann, C., and Dumas, P. (2001). Crystal structures of coaxially stacked kissing complexes of the HIV-1 RNA dimerization initiation site. *Nature Struct. Biol.*, **8**, 1064–1068.
- Ennifar, E., Yusupov, M., Walter, P., Marquet, R., Ehresmann, B., Ehresmann, C., and Dumas, P. (1999). The crystal structure of the dimerization initiation site of genomic HIV-1 RNA reveals an extended duplex with two adenine bulges. *Structure*, **7**, 1439–1449.
- Ennifar, E., Nikulin, A., Tishchenko, S., Serganov, A., Nevskaya, N., Garber, M., Ehresmann, B., Ehresmann, C., Nikonov, S., and Dumas, P. (2000). The crystal structure of UUCG tetraloop. *J. Mol. Biol.*, **304**, 35–42.
- Feig, A. L., Scott, W. G., and Uhlenbeck, O. C. (1998). Inhibition of the hammerhead ribozyme cleavage reaction by site-specific binding of Tb. *Science*, **279**, 81–84.
- Ferré-D'Amaré, A. R., Zhou, K., and Doudna, J. A. (1998). Crystal structure of a hepatitis delta virus ribozyme. *Nature*, **395**, 567–574.
- Ferre-D'Amare, A. R., Zhou, K., and Doudna, J. A. (1998). A general module for RNA crystallization. *J. Mol. Biol.*, **279**, 621–631.
- François, B., Szychowski, J., Adhikari, S.S., Pachamuthu, K., Swayze, E. E., Griffey, R. H., Migawa, M. T., Westhof, E., and Hanessian, S. (2004). Antibacterial aminoglycosides with a modified mode of binding to the ribosomal-RNA decoding site. *Angew. Chem. Int. Ed.*, **43**, 6735–6738.
- Gilbert, S. D., Mediatore, S. J., and Batey, R. T. (2006). Modified pyrimidines specifically bind the purine riboswitch. *J. Amer. Chem. Soc.*, **351**, 371–382.
- Golden, B. L. (2007). Preparation and crystallization of RNA. *Methods Mol. Biol.*, **363**, 239–257.
- Golden, B. L., Kim, H., and Chase, E. (2005). Crystal structure of a phage Twort group 1 ribozyme-product complex. *Nature Struct. Mol. Biol.*, **12**, 82–89.
- Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N., and Altman, S. (1983). The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, **35**, 849–857.
- Guo, F., Gooding, A. R., and Cech, T. R. (2004). Structure of the *Tetrahymena* ribozyme: base triple sandwich and metal ion at the active site. *Mol. Cell*, **16**, 351–362.
- Hanessian, S., Szychowski, J., Adhikari, S. S., Vasquez, G., Kandasamy, P., Swayze, E. E., Migawa, M. T., Ranken, R., François, B., Wirmer-Bartoschek, J., Kondo, J., and Westhof, E. (2007). Structure-based design, synthesis, and A-site rRNA cocrystal complexes of functionally novel aminoglycoside antibiotics: C2'' ether analogues of paromomycin. *J. Med. Chem.*, **50**, 2352–2369.
- Hansen, J. L., Ippolito, J. A., Ban, N., Nissen, P., Moore, P. B., and Steitz, T. A. (2002). The structures of four macrolide antibiotics bound to the large ribosomal subunit. *Mol. Cell*, **10**, 117–128.
- Heus, H. A. and Pardi, A. (1991). Structural features that give rise to the unusual stability of RNA hairpins containing GNRA loops. *Science*, **253**, 191–194.
- Höbartner, C., Rieder, R., Kreutz, C., Puffer, B., Lang, L., Polonskaia, A., Serganov, A., and Micura, R. (2005). Syntheses of RNAs with up to 100 nucleotides containing site-specific 2'-methylseleno labels for use in X-ray crystallography. *J. Amer. Chem. Soc.*, **127**, 12035–12045.
- Holbrook, S. R., Cheong, C., Tinoco, I., and Kim, S.-H. (1991). Crystal structure of an RNA double helix incorporating a track of non-Watson-Crick base pairs. *Nature*, **353**, 579–581.
- Ippolito, J. A. and Steitz, T. A. (2000). The structure of the HIV-1 RRE high affinity rev binding site at 1.6 Å resolution. *J. Mol. Biol.*, **295**, 711–717.
- Jovine, L., Djordjevic, S., and Rhodes, D. (2000). The crystal structure of yeast phenylalanine tRNA at 2.0 Å resolution: cleavage by Mg²⁺ in 15-year old crystals. *J. Mol. Biol.*, **301**, 401–414.
- Ke, A. and Doudna, J. A. (2004). Crystallization of RNA and RNA-protein complexes. *Methods*, **34**, 408–414.
- Kim, S.-H., Suddath, F. L., Quigley, G. J., McPherson, A., Sussman, J. L., Wang, A. H.-J., Seeman, N. C., and Rich, A. (1974). Three-dimensional tertiary structure of yeast phenylalanine transfer RNA. *Science*, **185**, 435–440.
- Klosterman, P. S., Shah, S. A., and Steitz, T. A. (1999). Crystal structures of two plasmid copy control related rna duplexes: an 18 base pair duplex at 1.20 Å resolution and a 19 base pair duplex at 1.55 Å resolution. *Biochemistry*, **38**, 14784–14792.
- Korostelev, A., Trakhanov, S., Laurberg, M., and Noller H. F. (2006). Crystal structure of a 70S ribosome-tRNA complex reveals functional interactions and rearrangements. *Cell*, **126**, 1065–1077.

- Kundrot, C. (1997). Preparation and Crystallization of RNA: A Sparse Matrix Approach. *Meth. Enzymol.*, **276**, 143–156.
- Lee, J. C. and Gutell R. R. (2004). Diversity of base-pair conformations and their occurrence in rRNA structure and RNA structural motifs. *J. Mol. Biol.* **344**, 1225–1249.
- Leontis, N. B., Altman, R. B., Berman, H. M., Brenner, S. E., Brown, J. W., Engelke, D. R., Harvey, S. C., Holbrook, S. R., Jossinet, F., Lewis, S. E., Major, F., Mathews, D. H., Richardson, J. S., Williamson, J. R., and Westhof, E. (2006). The RNA Ontology Consortium: An open invitation to the RNA community. *RNA*, **12**, 533–541.
- Leontis, N. B., Stombaugh, J., and Westhof, E. (2002). The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res.*, **30**, 3497–3531.
- Leontis, N. B. and Westhof, E. (2001). Geometric nomenclature and classification of RNA base pairs. *RNA*, **7**, 499–512.
- Lescoute, A., Leontis, N. B., Massire C. and Westhof, E. (2005). Recurrent structural RNA motifs, isostericity matrices and sequence alignments. *Nucleic Acids Res.*, **33**, 2395–2409.
- Lescoute, A. and Westhof, E. (2006). Topology of three-way junctions in folded RNAs. *RNA*, **12**, 83–93.
- Lilley, D. M. J. (1999). Structure, folding and catalysis of the small nucleolytic ribozymes. *Curr. Opin. Struct. Biol.*, **9**, 330–338.
- Ma, J.-B., Ye, D., and Patel, D. J. (2004). Structural basis for overhang-specific small interfering RNA recognition by the PAZ domain. *Nature*, **429**, 318–322.
- Ma, J.-B., Yan, Y.-R., Meister, G., Pei, Y., Tuschl, T., and Patel, D. J. (2005). Structural basis of guide RNA by the *A. fulgidus* Piwi protein. *Nature*, **434**, 666–670.
- Miller, J. L. and Kollman, P. A. (1997). Theoretical studies of an exceptionally stable RNA tetraloop: observation of convergence from an incorrect NMR structure to the correct one using unrestrained molecular dynamics. *J. Mol. Biol.*, **270**, 436–450.
- Mooers, B. H. M., Logue, J. S., and Berglund, J. A. (2005). The structural basis of myotonic dystrophy from the crystal structure of CUG repeats. *Proc. Natl. Acad. Sci. USA*, **102**, 16626–16631.
- Murray, J. B., Meroueh, S. O., Russell, R. J., Lentzen, G., Haddad, J., and Mobashery, S. (2006). Interactions of designer antibiotics and the bacterial ribosomal aminoacyl-tRNA site. *Chem. Biol.*, **13**, 129–138.
- Murray, J. B., Terwey, D. P., Maloney, L., Karpeisky, A., Usman, N., Beigelman, L., and Scott, W. G. (1998). The structural basis of hammerhead ribozyme self-cleavage. *Cell*, **92**, 665–673.
- Murray, J. B., Szöke, H., Szöke, A., and Scott, W. G. (2000). Capture and visualization of a catalytic RNA enzyme-product complex using crystal lattice trapping and X-ray holographic reconstruction. *Mol. Cell*, **5**, 279–287.
- Nissen, P., Ippolito, J. A., Ban, N., Moore, P. B., and Steitz, T. A. (2001). RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. *Proc. Natl. Acad. Sci. USA*, **98**, 4899–4903.
- Ogle, J. M., Broderson, D. E., Clemons, W. M., Tarry, M. J., Carter, A. P., and Ramakrishnan, V. (2001). Recognition of cognate transfer RNA by the 30S ribosomal subunit. *Science*, **292**, 897–902.
- Pallan, P. S., Marshall, W. S., Harp, J., Jewett III, F. C., Wawrzak, Z., Brown II, B. A., Rich, A., and Egli M. (2005). Crystal structure of a luteoviral RNA pseudoknot and model for a minimal ribosomal frameshifting motif. *Biochemistry*, **44**, 11315–11322.
- Parker, J. S., Rose M. S., and Barford D. (2005). Structural insights into mRNA recognition from a PIWI domain-siRNA guide complex. *Nature*, **434**, 663–670.
- Peterson, R. D. and Feigon, J. (1996). Structural change in Rev responsive element RNA of HIV-1 on binding Rev peptide. *J. Mol. Biol.*, **264**, 863–877.
- Pley, H. W., Flaherty, K. M., and McKay, D. B. (1994). Three-dimensional structure of a hammerhead ribozyme. *Nature*, **372**, 68–74.
- Poehlsgaard, J. and Douthwaite, S. (2005). The bacterial ribosome as a target for antibiotics. *Nature Rev. Microbiol.*, **3**, 870–881.
- Portmann, S., Usman, N., and Egli, M. (1995). The crystal structure of r(CCCCGGGG) in two distinct lattices. *Biochemistry*, **34**, 7569–7575.
- Price, S. R., Ito, N., Oubridge, C., Avis, J. M., and Nagai, K. (1995). Crystallization of RNA-protein complexes. I. Methods for the large-scale preparation of RNA suitable for crystallographic studies. *J. Mol. Biol.*, **249**, 398–408.
- Rife, J. P., Stallings, S. C., Correll, C. C., Dallas, A., Steitz, T. A., and Moore, P. B. (1999). Comparison of the crystal and solution structures of two RNA oligonucleotides. *Biophys. J.*, **76**, 65–75.
- Robertus, J. D., Ladner, J. E., Finch, J. T., Rhodes, D., Brown, R. S., Clark, B. F. C., and Klug, A. (1974). Structure of yeast phenylalanine tRNA at 3 Å resolution. *Nature*, **250**, 546–551.
- Rosenberg, J. M., Seeman, N. C., Day, R. O., and Rich, A. (1976). RNA double-helical fragments at atomic resolution. II. The crystal structure of sodium guanylyl-3',5'-cytidine nonahydrate. *J. Mol. Biol.*, **104**, 145–167.
- Rupert, P. B. and Ferré-D'Amaré, A. R. (2001). Crystal structure of a hairpin ribozyme-inhibitor complex with implications for catalysis. *Nature*, **410**, 780–786.
- Sarkhel, S., Rich, A., and Egli, M. (2003). Water-nucleobase “stacking”: H- π and lone pair- π interactions in the atomic resolution crystal structure of an RNA pseudoknot. *J. Amer. Chem. Soc.*, **125**, 8998–8999.

- Schindelin, H., Zhang, M., Bald, R., Fürste, J.-P., Erdmann, V. A., and Heinemann, U. (1995). Crystal structure of an RNA dodecamer containing the Escherichia coli Shine-Dalgarno sequence. *J. Mol. Biol.*, **249**, 595–603.
- Schlunzen, F., Tocilj, A., Zarivach, R., Harms, J., Gluhmann, M., Janell, D., Bashan, A., Bartles, H., Agmon, I., Franceschi, F., and Yonath, A. (2000). Structure of functionally activated small ribosomal subunit at 3.3 Å resolution. *Cell*, **102**, 615–623.
- Schneider, B., Morávek, Z., and Berman, H. M. (2004). RNA conformational classes. *Nucleic Acids Res.*, **32**, 1666–1677.
- Schuwirth, B. S., Borovinskaya, M. A., Hau, C. W., Zhang, W., Vila-Sanjurjo, A., Holton J. M., and Doudna Cate, J. H. (2005). Structures of the bacterial ribosome at 3.5 Å resolution. *Science*, **310**, 827–833.
- Schwalbe, H. Buck, J., Fürtig, B., Noeske, J., and Wöhnert, J. (2007). Structures of RNA switches: insight into molecular recognition and tertiary structure. *Angew. Chem. Int. Ed.*, **46**, 2–10.
- Scott, W. G., Finch, J. T., and Klug, A. (1995). The crystal structure of an all-RNA hammerhead ribozyme: a proposed mechanism for RNA catalytic cleavage. *Cell*, **81**, 991–1002.
- Seeman, N. C., Rosenberg, J. M., Suddath, F. L., Kim, J. J. P., and Rich, A. (1976). RNA double-helical fragments at atomic resolution. I. The crystal and molecular structure of sodium adenylyl-3',5'-uridine hexahydrate. *J. Mol. Biol.*, **104**, 109–144.
- Selmer, M., Dunham, C. M., Murphy IV, F. V., Weixlbaumer, A., Petry, S., Kelley, A. C., Weir, J. R., and Ramakrishnan, V. (2006). Structure of the 70S ribosome complexed with mRNA and tRNA. *Science*, **313**, 1935–1941.
- Serganov, A., Polonskaia, A., Phan, A. T., Breaker, R. R., and Patel D.J. (2006). Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch. *Nature*, **441**, 1067–1171.
- Serganov, A., Yuan, Y. R., Pitovskaya, O., Polonskaia, A., Malinina, L., Phan, A. T., Hobartner, C., Micura, R., Breaker, R. R., and Patel D.J. (2004). Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs. *Chem. Biol.*, **11**, 1729–1741.
- Shi, H. and Moore, P. B. (2000). The crystal structure of yeast phenylalanine tRNA at 1.93 Å resolution: a classic structure revisited. *RNA*, **6**, 1091–1105.
- Strobel, S. A. and Ryder, S. P. (2001). Biological catalysis. The hairpin's turn. *Nature*, **410**, 761–763.
- Sundaralingam, M. and Pan, B. (2002). Hydrogen and hydration of DNA and RNA oligonucleotides. *Biophys. Chem.*, **95**, 273–282.
- Sykes, M. T. and Levitt, M. (2005). Describing RNA structure by libraries of clustered nucleotide doublets. *J. Mol. Biol.*, **351**, 26–38.
- Tereshko, V., Skripkin, E., and Patel, D. J. (2003). Encapsulating streptomycin within a small 40-mer RNA. *Chem. Biol.*, **10**, 175–187.
- Theimer, C. A., Blois, C. A., and Feigon, J. (2005). Structure of the human telomerase RNA pseudoknot reveals conserved tertiary interactions essential for function. *Mol. Cell.*, **17**, 671–682.
- Thore, S., Leibundgut, M., and Ban, N. (2006). Structure of the eukaryotic thiamine pyrophosphate riboswitch with its regulatory ligand. *Science*, **312**, 1208–1211.
- Tocilj, A., Schlunzen, F., Janell, D., Gluhmann, M., Hansen, H. A. S., Harms, J., Bashan, A., Bartles, H., Agmon, I., Franceschi, F., and Yonath, A. (1999). The small ribosomal subunit from Thermus thermophilus at 4.5 Å resolution: pattern fittings and the identification of a functional site. *Proc. Natl. Acad. Sci. USA*, **96**, 14252–14257.
- Tu, D., Blaha, G., Moore, P. B., and Steitz, T. A. (2005). Structures of MLSBK antibiotics bound to mutated large ribosomal subunits provide a structural explanation for resistance. *Cell*, **121**, 257–270.
- Varani, G. and McClain, W. H. (2000). The G × U wobble base pair. A fundamental building block of RNA structure crucial to RNA function in diverse biological systems. *EMBO Reports*, **1**, 18–23.
- Vicens Q. and Cech, T. R. (2005). Atomic level architecture of group I introns revealed. *Trends Biochem. Sci.*, **31**, 41–51.
- Vicens, Q. and Westhof, E. (2001). Crystal structure of paromomycin docked into the eubacterial ribosomal decoding A site. *Structure*, **9**, 647–658.
- Wedekind, J. E. and McKay, D. B. (2000). Purification, crystallization, and X-ray diffraction analysis of small ribozymes. *Methods in Enzymology*, **317**, 149–168.
- Westhof, E. and Sundaralingam, M. (1986). Restrained refinement of the monoclinic form of yeast phenylalanine transfer RNA. Temperature factors and dynamics, coordinated waters, and base-pair propeller twist angles. *Biochemistry*, **25**, 4868–4878.
- Westhof, E., Dumas, P. and Moras, D. (1988). Restrained refinement of two crystalline forms of yeast aspartic acid and phenylalanine transfer RNA crystals. *Acta Crystallogr.*, **A44**, 112–123.
- Wild, K., Weichenrieder, O., Leonard, G. A., and Cusack, S. (1999). The 2 Å structure of helix 6 of the human signal recognition particle RNA. *Structure*, **7**, 1345–1352.
- Wimberly, B. T., Brodersen, D. E., Clemons, W. M., Morgan-Warren, R. J., Carter, A. P., Vonrein, C., Hartsch, T., and Ramakrishnan, V. (2000). Structure of the 30S ribosomal subunit. *Nature*, **407**, 327–339.
- Winkler, W. C., Nahvi, A., Roth, A., Collins, J. A., and Breaker, R. R. (2004). Control of gene expression by a natural metabolite-responsive ribozyme. *Nature*, **428**, 261–286.

- Wu, H., Henras, A., Chanfreau, G., and Feigon, J. (2004). Structural basis for recognition of the AGNN tetraloop RNA fold by the double-stranded RNA-binding domain of Rnt1p RNase III. *Proc. Natl. Acad. Sci. USA*, **101**, 8307–8312.
- Yusupov, M. M., Yusupova, G. Z., Baucom, A., Lieberman, K., Earnest, T. N. Cate, J. H. D., and Noller, H. F. (2001). Crystal structure of the ribosome at 5.5 Å resolution. *Science*, **292**, 883–896.

Further Reading

RNA Structures

- Batey, R. T., Rambo, R. P., and Doudna, J. A. (1999). Tertiary motifs in RNA structure and folding. *Angew. Chem. Int. Edit.*, **38**, 2326–2343.
- Fürtig, B., Richter, C., Wühnert, J., and Schwalbe, H. (2003). NMR Spectroscopy of RNA. *ChemBioChem.*, **4**, 10–36.
- Latham, M. P., Brown, D. J., McCallum, S. A., and Pardi, A. (2005). NMR methods for studying the structure and dynamics of RNA. *Chem. Biochem.*, **6**, 1492–1505.
- Leontis, N. B. and Westhof, E. (2003). Analysis of RNA motifs. *Curr. Opin. Struct. Biol.*, **13**, 300–308.
- Shen, L. X., Cai, Z., and Tinoco, I. (1995). RNA structure at high resolution. *FASEB Journal*, **9**, 1023–1033.
- Westhof, E. (1999). in *Oxford Handbook of Nucleic Acid Structure* (Neidle, S., Ed.), p. 533, Oxford University Press, Oxford.

tRNA

- Arnez, J. G. and Moras, D. (1999). in *Oxford Handbook of Nucleic Acid Structure* (Neidle, S., Ed.), p. 603, Oxford University Press, Oxford.
- Clark, B. F. (2006). The crystal structure of tRNA. *J. Biosci.*, **31**, 453–457.
- Kim, S.-H. (1981). In *Topics in Nucleic Acid Structure* (Neidle, S., Ed.), p. 83, Macmillan Press, London.

Ribozymes

- Butcher, S. E. (2001). Structure and function of the small ribozymes. *Curr. Opin. Struct. Biol.*, **11**, 315–320.
- Doherty, E. A. and Doudna, J. A. (2001). Ribozyme structures and mechanisms. *Ann. Rev. Biophys.*, **30**, 457–475.
- Ferré-D'Amaré, A. R. and Doudna, J. A. (1999). RNA folds: insights from recent crystal structures. *Ann. Rev. Biophys.*, **28**, 57–73.
- Lilley, D. M. J. (2005). Structure, folding and mechanisms of ribozymes. *Curr. Opin. Struct. Biol.*, **15**, 313–323.
- Scott, W. G. (2007). Ribozymes. *Curr. Opin. Struct. Biol.*, **17**, in press.

Ribosome Structure

- Cech, T. R. (2000). Structural biology. The ribosome is a ribozyme. *Science*, **289**, 878–879.
- Maguire, B. A. and Zimmermann, R. A. (2001). The ribosome in focus. *Cell*, **104**, 813–816.
- Moore, P. B. (2001). The ribosome at atomic resolution. *Biochemistry*, **40**, 3243–3250.
- Moore, P. B. and Steitz, T. A. (2003). The structural basis of large ribosomal subunit function. *Ann. Rev. Biochem.*, **72**, 813–850.
- Noller, H. F. (2005). RNA structure: reading the ribosome. *Science*, **309**, 1508–1514.
- Pennisi, E. (1999). The race to the ribosome structure. *Science*, **285**, 2048–2051.

Riboswitches

- Schwalbe, H., Buck, J., Fürtig, B., Noeske, J., and Wöhnert, J. (2007). Structures of RNA switches: insight into molecular recognition and tertiary structure. *Angew. Chem. Int. Edit.*, **46**, 2–10.
- Tucker, B. J. and Breaker, R. R. (2005). Riboswitches as versatile gene control elements. *Curr. Opin. Struct. Biol.*, **15**, 342–348.

RNA-Drug Complexes

- Gallego, J. and Varani, G. (2001). Targeting RNA with small-molecule drugs: therapeutic promise and chemical challenges. *Acc. Chem. Res.*, **34**, 836–843.
- Hermann, T. (2005). Drugs targeting the ribosome. *Curr. Opin. Struct. Biol.*, **15**, 355–366.
- Vicens, Q and Westhof, E. (2003). RNA as a drug target: the case of aminoglycosides. *ChemBiochem.*, **4**, 1018–1023.
- Wirmer, J. and Westhof, E. (2006). Molecular contacts between antibiotics and the 30S ribosomal particle. *Methods Enzymol.*, **415**, 180–202.

Principles of Protein-DNA Recognition

7.1 Introduction

The sequence information within DNA exists to be replicated, transcribed, and translated into gene products, depending on the nature of the cell type, its progression through the cell cycle and its response to external stimuli and factors. The processes of packaging DNA in the cell, regulating the expression of individual genes, and then reading/ translating the encoded sequences, requires the interplay of large numbers of nonspecific and specific proteins. Similarly the responses to DNA damage involve the concerted interplay of a number of proteins, to first recognize damage, then to excise it, and finally to repair the DNA. It is thus unsurprising that the study of DNA-protein interactions has progressively developed, initially as a major area in its own right, and now into a number of more specialized areas, each with a large body of structural knowledge underpinning the biology and biophysics. This chapter concentrates on the underlying principles of recognition as found by structural methods (principally X-ray crystallography and NMR), and how DNA structure itself copes with a wide range of protein and functional requirements. The reader is directed to the extensive primary literature for detailed descriptions of individual structures.

Structural information is now available on a large (and still increasing) number of distinct DNA-protein complexes, from a wide range of eukaryotic and prokaryotic sources. As at June 2007, the PDB and NDB have 1789 protein-DNA structures; in 1997 there were just 241 structures. Studies of the proteins themselves rarely provide sufficient insight into the processes of recognition, and are not discussed in any detail in this book. This still relatively small number of known protein-DNA structures pales into insignificance compared to the total encoded by individual genomes. The determination of the human genome sequence in 2001 has enabled reliable estimates to be made for the numbers of genes with particular functions in just this one genome. One estimate, taken from the Transcription Factor Database www.transcriptionfactor.org (Kummerfeld and Teichmann, 2006), uses the longest transcript per gene as the single representative protein for each gene, and estimates that there are 1489 transcription factors out of a total of 22,983 genes in the human genome. We are currently seeing a major effort in large-scale high-throughput structural studies (structural genomics), that will in time reducing this disparity between numbers of known and unknown structures. However major increases in our understanding of DNA-protein recognition and processing will not be obtained solely from complexes with single proteins, but increasingly from studies of functionally relevant multiprotein complexes, as we

have seen in the case of the ribosome for translation, and the polymerase machinery for transcription.

DNA-binding proteins can be conveniently categorized in both functional and structural terms into a number of families. The principal functional groupings are:

- Regulatory proteins. These mostly bind to highly specific sequences of duplex DNA, in order to control the transcription of a particular gene, or bind to particular signal sequences such as 5'-TATA in order to more generally initiate transcription.
- DNA cleavage proteins (nucleases). Some such as DNase I have relatively little sequence specificity (although they may have some DNA structural selectivity). Others, such as the restriction enzymes, are highly specific for particular sequences.
- Repair proteins that respond to various types of damage to DNA by recognizing the lesion itself, then excising the damaged DNA, and/or joining together breaks in damaged DNA.
- Proteins that resolve topological problems in DNA by unravelling or unwinding DNA prior to replication. The DNA topoisomerases, of importance as cancer therapeutic targets, have been extensively studied.
- Structural proteins that maintain the integrity of folded or packaged DNA, for example, histones in chromatin.
- Processing proteins, typified by DNA and RNA polymerases, that use DNA as a template for further nucleic acid synthesis. Sequence specificity is absolutely not required for DNA recognition, rather a need to recognize a particular type of DNA duplex.

This diversity of functions shown by DNA-binding proteins is in striking contrast to the relatively few ways in which DNA structure is “read,” even though there is also very wide variety in the structure of the proteins themselves. Direct reading of a DNA sequence generally occurs via the hydrogen-bonding edges of the bases (Seeman, Rosenberg, and Rich, 1976), as described in Chap. 5 for small molecules; therefore features are required of a protein that enable the bases to be accessed through either major or minor grooves. Figure 7.1 shows that the B-DNA major groove is the richer of the two grooves of duplex DNA, both in information content per se, and in its ability to facilitate discrimination between different DNA sequences, which is essential if the appropriate genes are to be transcribed. Thus, the major groove is generally the site of direct information readout. Nonetheless, the minor groove is an important target for some regulatory and structural proteins, especially those that are able to deform DNA so that the minor groove becomes greatly expanded. Indirect readout can occur via the sugar-phosphate backbone or solely the phosphate groups. Especially when this is sequence-specific, it is the cumulative consequence of a number of individually nonspecific nonbonded interactions as compared to the clearly defined hydrogen-bonding interactions involved in direct recognition. The anionic nature of phosphate groups makes them frequent targets for basic side chains, with the resulting electrostatic interactions being significant contributors to overall protein-DNA binding.

We normally assume, quite rightly, that the base-pairing in B-DNA comprises solely Watson–Crick type, and therefore that proteins need to recognize Watson–Crick base edges. However this is probably not universally the case, since there is a small number of well-authenticated protein-DNA structures in which Hoogsteen

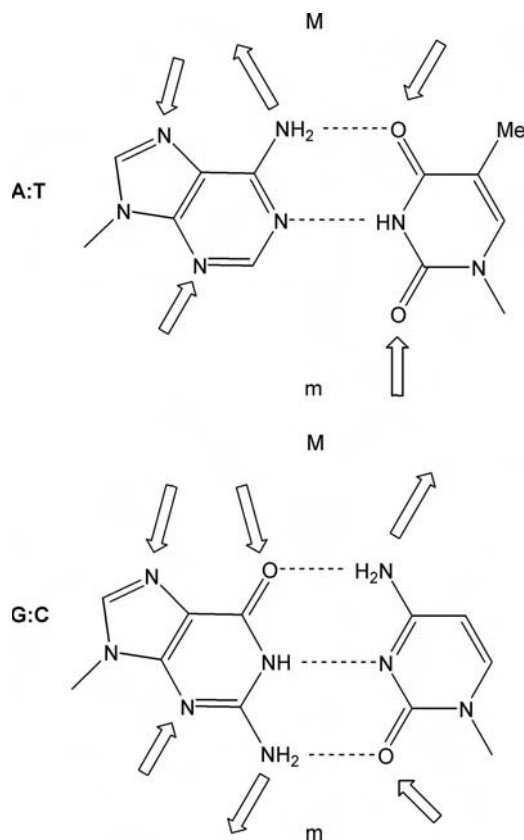


Figure 7.1 The pattern of hydrogen-bond donors and acceptors in Watson–Crick base pairs. The direction of the arrows indicates the direction of hydrogen-bonding donation.

base-pairs have been observed. One example is the high-resolution structure of the MAT α 2 homeodomain DNA complex (Aishima *et al.*, 2002), in which one A•T base pair was unequivocally assigned to be of Hoogsteen type, although an NMR study showed that it was not formed in the native DNA structure. It was concluded on the basis of the intermolecular contacts that the Hoogsteen-pairing was stabilized by the protein-binding. A more profound occurrence has been observed in the crystal structure of human DNA polymerase ϵ bound to a template primer DNA (Nair *et al.*, 2004, 2006), in which the incoming nucleotide triphosphate forces the adenine and guanines of the template at the active site into a *syn* conformation, and thus to form a Hoogsteen A•T base pair.

There are a limited number of ways by which proteins recognize duplex DNA (Harrison, 1991; Luscombe *et al.*, 2000). This is unsurprising given that the majority of protein-DNA complexes retain DNA in an overall B-type conformation, and as we have seen in earlier chapters, the deviations from ideality are rarely large. The principal recognition motifs found in DNA-binding proteins are:

- α -helix-turn- α -helix (the HTH motif)
- Zinc-finger and loop

- Zipper motifs
- β -sheets
- β -hairpins

These are described below.

7.2 Direct Protein-DNA Contacts

These interactions involve hydrogen bonds between amino-acid side-chains and the edges of the base pairs, involving their pattern of donors and acceptors molecules. The same principles of hydrogen-bonding apply as for small molecules binding to DNA (see Chap. 5), although the common contact areas involved in protein-DNA interfaces are invariably much larger (Nadassy, Wodak, and Janin, 1999). The protein backbone itself does sometimes make contact with bases or phosphates, but these are generally not determinants of specificity, rather they serve to enhance binding affinity. The majority of interactions involve O6 and/or N7 atoms of guanine bases forming hydrogen bonds with the charged ends of long flexible side-chains from the basic residues arginine or lysine, the amide residues glutamine and asparagine or the hydroxyl group of a serine (Table 7.1 and Figs. 7.2, 7.3). The recognition of one hydrogen-bonding site on a base, compared to two simultaneous sites on that base, actually involves only a small change in position of the amino-acid side chain. In general, the latter will be energetically preferred and formed if possible. Adenine bases are recognized via their N6 and/or N7 atoms, although this occurs less frequently than guanine recognition. Active pyrimidine recognition is also slightly less common than guanine recognition.

There is no general 1:1 amino acid:DNA base correspondence, and recognition can sometimes occur in a wide variety of ways in addition to the simple mono- or bidentate ones. Table 7.1 also shows that hydrogen-bonding to both partners simultaneously in Watson–Crick base pairs is relatively uncommon, not least since binding to one usually results in sufficient discrimination for sequence selectivity to take place. Some arrangements for side-chains bridging between two bases in a base pair are shown in Fig. 7.4a,b. Hydrogen-bonding can also occur to two consecutive bases on the same strand (Fig. 7.5). This arrangement introduces sequence selectivity at the dinucleotide level. For example, the TG sequence shown in Fig. 7.5 is recognized by the lysine side-chain through both carbonyl groups of the thymine and guanine bases. Altering this to TA or TC would not

Table 7.1 Distribution of Amino Acid-Base Interactions in Protein-DNA Structures, Adapted from (Luscombe, Laskowski, and Thornton, 2001), Using a Data Set of 129 Structures. Only those Amino Acids that Participate in a Significant Number of DNA Base Interactions are Listed

	Guanine	Cytosine	Adenine	Thymine
Arginine	98	8	19	24
Lysine	30	6	4	9
Serine	12	2	1	3
Asparagine	7	10	18	7
Glutamine	6	2	16	2
Glutamate	1	10	1	0

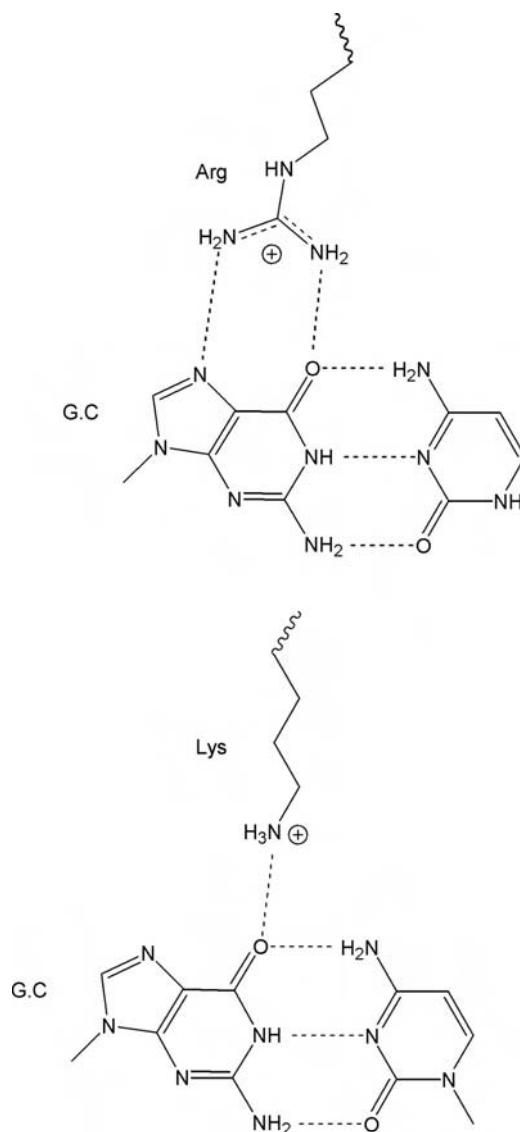


Figure 7.2 Possible patterns of hydrogen-bonding recognition of G•C base pairs by arginine and lysine side-chains.

put two carbonyl groups in the same correct position for interaction with lysine. Similar arguments can be made for other dinucleotide steps involved in bidentate recognition with side-chains.

Sometimes a water molecule (or molecules) participates in such a hydrogen-bonded bridge, or itself bridges between a base and a side chain. Such water involvement has been found, for example, in the *E. Coli trp* repressor/operator complex (Otwinowski *et al.*, 1988), with only two (relatively unimportant) direct (guanine... arginine) contacts, yet three that have water-mediated contacts. The crystal structure of the excisionase-DNA complex from bacteriophage lambda

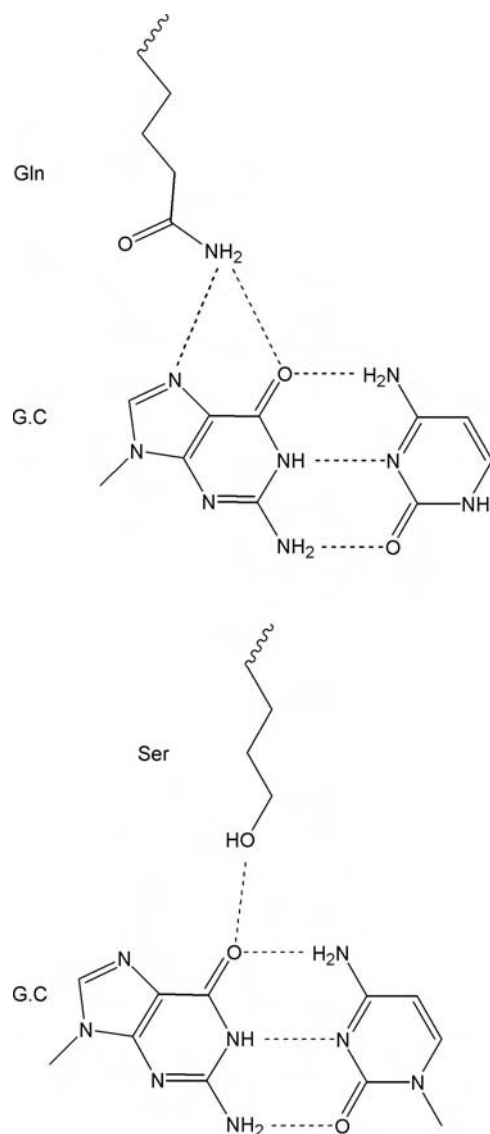


Figure 7.3 Possible patterns of hydrogen-bonding recognition of G•C base pairs by glutamine and serine side-chains.

(Sam *et al.*, 2004) shows an unusual winged-helix motif (α -helix + β -hairpin wing) in the excisionase protein interacting in the B-DNA major groove via an extensive network of water-mediated contacts. A further example of the important role that water molecules can play, has been seen in the crystal structure of the bacteriophage λ repressor-operator complex (Beamer and Pabo, 1992) where the hydroxyl group of serine-45 interacts directly with O6 and N7 of a guanine. The backbone carbonyl oxygen atom of this serine interacts via a water molecule with the N4 atom of the complementary cytosine base and through the same water molecule to O4 of the adjacent thymine.

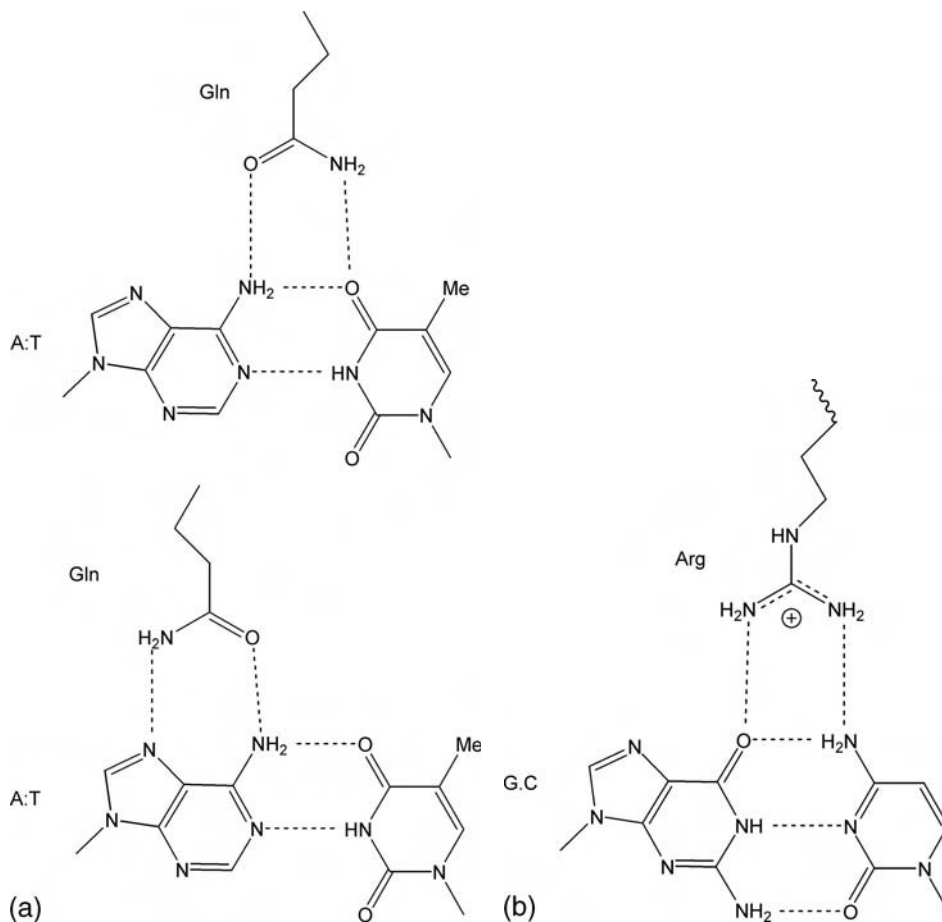


Figure 7.4 (a) Two possible patterns of hydrogen-bonding recognition of A•T base pairs by glutamine side-chains. (b) The pattern of arginine recognition of both bases in a G•C base pair.

It has not been possible to establish a consistent pattern of preference between a particular base and an amino-acid residue from surveys of known structures. This lack of a general hydrogen-bonding recognition pattern (although there are some involving very closely related proteins), reflects a number of factors (Matthews, 1988; Mandel-Gutfreund, Schueler, and Margalit, 1995; Pabo and Nekludova, 2000):

- The effect of differences in DNA structure in various complexes
- The different types of protein motif involved in DNA recognition
- Other recognition factors that can play significant roles in defining the recognition of particular sequences, notably the ability of thymine methyl groups to form close van der Waals attractive interactions with hydrophobic side chains
- In general any one base can be recognized by several amino acids

An example of the role of thymine methyl groups is in the engrailed homeodomain structure, where there is a close contact between isoleucine-47 and one such methyl

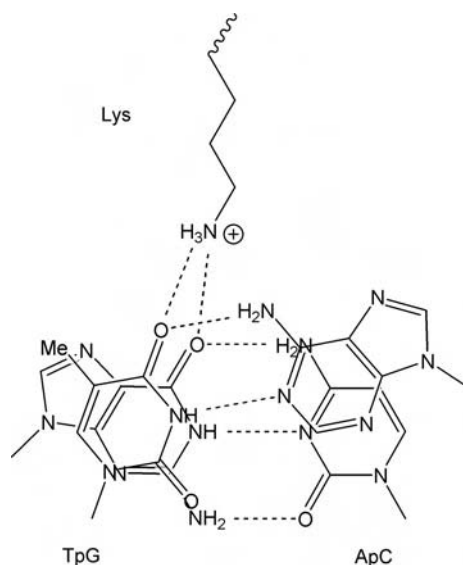


Figure 7.5 A possible recognition mode of the sequence TpG by a lysine side-chain.

group from a thymine in the TAAT recognition site (Kissinger *et al.*, 1990). At least as important roles are played in stabilizing a protein-DNA complex as a whole by numerous nonspecific (at least in hydrogen-bonding terms) protein-backbone and other interactions. Thus the flexibility of amino-acid side-chains ensures that optimal interactions are made with all elements of a DNA structure. These cumulatively enable the recognition helix to be oriented in the major groove in a manner that is distinctive for a particular protein. Such generalized hydrophobic and electrostatic interactions may also play a more active recognition role, by means of indirect readout, sensing the DNA structure and flexibility specified by a particular sequence. In view of all of these factors it is then unsurprising that there is wide variation in the number and importance of direct readout interactions that have been observed in different HTH complexes. The major role that can be occasionally played by indirect readout is strikingly demonstrated in the structure of the *trp* repressor/operator complex (Otwinowski *et al.*, 1988). This complex has a large number of contacts between side-chains and the phosphate groups of the operator DNA, yet very surprisingly has no functionally significant direct readout base-amino acid interactions at all. In spite of this, the cumulative effect of the indirect readout contacts made by the *trp* repressor is to effectively read the particular structural details of its DNA operator sequence (Haran, Joachimiak, and Sigler, 1992).

Readout, both direct and indirect, ensures that key residues on both protein and DNA, are effectively recognized. The functional importance of such residues and hence the protein-DNA contacts involved, can be evaluated by mutagenesis experiments, as well as by surveying their occurrence within a particular family of proteins and their consensus operator sequences. For example, residues tryptophan-48, phenylalanine-49, asparagine-51 and arginine-53 occur in every member of the large eukaryotic homeodomain family. The crystal structure of the engrailed homeodomain-DNA complex (Kissinger *et al.*, 1990) shows that the first two of these residues play critical roles in preserving the hydrophobic core

of the recognition helix. The other two directly interact with base and phosphate groups respectively.

7.3 Major-Groove Interactions – the α -Helix as the Recognition Element

A protein α -helix can fit snugly only into the major groove rather than the minor groove of a B-like DNA duplex and therefore act directly as the recognition element. This element was first observed in the bacterial *cro* repressor protein (Anderson *et al.*, 1981) as part of the helix-turn-helix (HTH) motif. This HTH domain pattern has subsequently been found in the crystal and solution NMR structures of a large number of prokaryotic and eukaryotic transcription regulatory and related proteins, and by comparative sequence analysis in many others (Table 7.2). The HTH motif consists of ~20 amino-acid residues with residues 1–7 forming the first α -helix and residues 12–20 the second α -helix. These are linked by a short turn, so that the two helices are inclined at 120° to each other. The second helix is the recognition one, which makes most of the specific contacts to DNA and lies in the major groove of a B-form duplex. There are characteristic hydrophobic residues at positions 4, 8, 10, 16, and 18 of the recognition helix, which help to form the overall hydrophobic core of the protein. The N-terminal region of homeodomains typically binds in the DNA minor-groove. The resulting side-chain interactions with minor-groove base edges are important contributors to overall homeodomain binding. The basic side-chain of, in particular, arginine closely mimics the behavior of basic side-chains in small molecules complexed to the DNA minor groove.

The HTH motif does not exist by itself, but is held together in a variety of ways. Thus, the eukaryotic homeodomain protein *engrailed* (Kissinger *et al.*, 1990; Wolberger *et al.*, 1991) has a three-helix bundle (Fig. 7.6), with the HTH motif itself occurring after the first helix, whereas the phage λ (Beamer and Pabo, 1992) and 434 repressor (Aggarwal *et al.*, 1988) structures are five-helix bundles with the HTH motif comprising helices 2 and 3 (Fig. 7.7). An important difference between the bacterial and eukaryotic HTH proteins is that the former generally bind to their target sites as dimers whereas the latter bind as monomers. This difference is reflected in the structures of the proteins themselves, with some aspects of the non-HTH domains being responsible for dimerisation. The features of these structures as seen in the crystalline state, are largely preserved in solution, although side-chain orientations are unsurprisingly, not always identical (Neri, Billeter, and Wüthrich, 1992; Fraenkel and Pabo, 1998).

Many homeodomains act at a DNA locus together with a second homeodomain, which can be binding in tandem at a neighboring site. This has been studied in detail

Table 7.2 Selected HTH Crystal and NMR Structures

Protein	PDB ID code	NDB ID code
434 repressor	3CRO	PDR001
λ repressor	4CRO	PDR003
<i>Lac</i> repressor	1EFA	PD0118
<i>Cro</i> repressor	3ORC	PD0022
<i>engrailed</i> homeodomain	3HDD	PD0016
<i>Antennapedia</i> homeodomain	1HOM	PDR055
MATa1/MAT α homeodomain	1AKH	PDR049
CAP	1CGP	PDR006

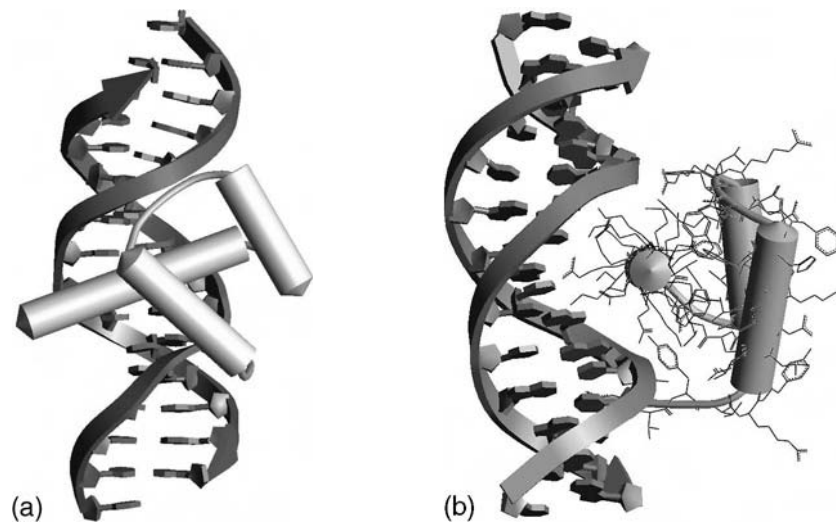


Figure 7.6 (a) The helix-turn-helix recognition of the bases in the DNA major groove in the *engrailed* homeodomain crystal structure (Kissinger *et al.*, 1990; Wolberger *et al.*, 1991). (b) A second view, now looking down the recognition helix of the homeodomain, and showing the side-chains extending into the major groove.



Figure 7.7 A view of the structure of the DNA-434 repressor complex (Aggarwal *et al.*, 1998). The recognition helices are shown oriented into the major groove.

for the MAT α 1 and MAT α 2 proteins, which regulate transcription in yeast. The crystal structure of the DNA complex with MAT α 2 alone (Wolberger *et al.*, 1991) shows a typical homeodomain, with a straight B-DNA helix. By contrast, the ternary complex (Li *et al.*, 1995) has the two protein units contacting each other through the C-terminus tail of the MAT α 2 domain. This occurs as a consequence of the DNA bending by 60°. A subsequent analogous crystal structure (Li *et al.*, 1998), this time using an A-tract sequence DNA, shows a very similar degree of bending, with the A-tract bent in the minor-groove direction. Remarkably, the minor-groove spine of hydration is preserved in both structures. Similar bending has been observed with a ternary complex of MAT α 2, the transcription factor MCM1 and a 26-base pair DNA sequence (Tan and Richmond, 1998). Other types of ternary complex can involve the two domains operating in close tandem to overlapping DNA sites, such as in the HoxB1-Pbx1 (Piper *et al.*, 1999) and Ubx-Exd (Passner *et al.*, 1999) heterodimer structures. Each protein in these complexes induces a small (10–11° in the former case) bend in the bound DNA, but since the binding sites are close and almost opposite each other, the net effect is an almost straight B-like helix. A general principle emerges from these studies, that when two protein domains bind on the same side of a DNA sequence, bending then ensues; when they are opposite, then there is no net bending.

The majority of HTH proteins have a number of direct interactions between the recognition helix and the major groove of the operator sequence DNA, which maintains a B-type conformation. However our knowledge of the detailed nature and extent of these interactions is critically dependent on reliable crystal and NMR structures. This caveat is illustrated by the redetermination at 2.2 Å resolution (Fraenkel *et al.*, 1998) of the original 2.8 Å *engrailed* homeodomain crystal structure (Kissinger *et al.*, 1990). The more recent of these two analyses confirms all major features of the complex, and also provides unequivocal information on the role of the important residue Gln50, which previously had been shown to interact only indirectly with a major-groove base edge. This role is confirmed in the higher resolution structure, which shows several water-mediated contacts between this residue and three (T4, T5, and G7) out of the seven bases in the d(TAATTAC) recognition site. The analysis also shows that there are important contacts between homeodomain side-chains and the phosphate backbone, some of which are mediated by water molecules. The case for the key role of water molecules in homeodomain recognition (Gehring *et al.*, 1994) has been reinforced by molecular dynamics simulations (Billeter *et al.*, 1996) and chemical modification studies (Labeots and Weiss, 1997). The simulations show that water molecules can have an appreciable residence time at the protein-DNA interface, sufficient for them to be mediating between DNA and amino acid side chains.

The widespread occurrence of the HTH motif has led to a number of studies to develop methods that can reliably locate the motif within genomic sequences using sequence data, either alone or together with a range of structural indicators such as the known structural information on HTH proteins, and electrostatic potential. The methods can achieve some success (see, e.g., Pellegrini-Calace and Thornton, 2005), when both sequence and structure are taken together.

7.4 Zinc-Finger Recognition Modes

Several zinc-containing motifs for sequence-specific DNA recognition are known, and their details have been revealed by crystallographic and NMR studies. They constitute

the largest family of DNA-binding motifs, and show considerable diversity in their architecture. A large group (and the first to be described, from the *Xenopus* transcription factor TFIIA) in this family is of the zinc-finger proteins that contain units of regularly spaced cysteine and histidine residues coordinated to a zinc ion (Klug and Rhodes, 1987; Laity, Lee, and Wright, 2001; Wolfe, Nekludova, and Pabo, 2000). Each finger consists of an anti-parallel β -sheet and an α -helix, held together by the zinc ion coordination (Fig. 7.8) (Klug and Rhodes, 1987). Zinc-finger proteins have at least two such units (“fingers”), with each recognizing a three base-pair site in the major groove of a B-DNA helix, as observed in the crystal structure of the Zif268 transcription factor (Pavletich and Pabo, 1991) (Table 7.3 and Fig. 7.9). Each finger makes direct contacts to the guanine-rich strand, with patterns of side-chain interactions from the α -helix involving arginine–guanine and histidine–guanine recognition that are very similar to those seen in HTH proteins Table 7.4.

A quite distinct zinc domain is involved in the structure of the yeast transcription activator GAL4 (Marmorstein *et al.*, 1992), which bind as a dimer. Each individual GAL4 molecule has two domains, a zinc-binding region and an α -helix, linked by an extended length of peptide (Fig. 7.10). The zinc-containing domain rests in the DNA major groove, where there is a pattern of direct interactions between

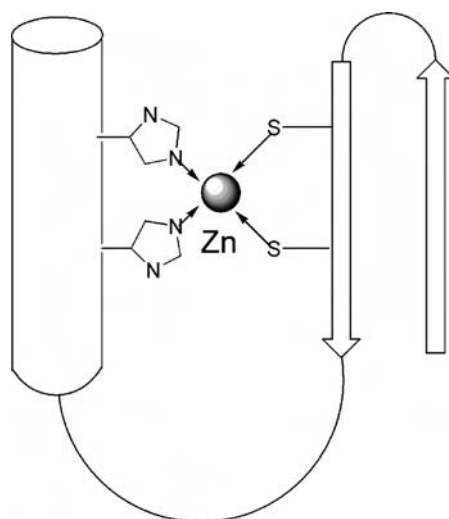


Figure 7.8 The zinc finger motif, showing the arrangement for a single finger, with the coordination to the zinc atom from histidine residues attached to the α -helix on the left, and the cysteine residues from the β -sheet on the right.

Table 7.3 Selected Zinc-Finger Protein-DNA Crystal Structures

Protein	PBD ID code	NDB ID code
Zif268	1ZAA	PDT006
GAL4	1D66	PDT003
Estrogen receptor	1HCQ	PDRC03
Retinoic acid receptor	2NLL	PDR021
TATA box-zinc finger	1G2D	PD0186

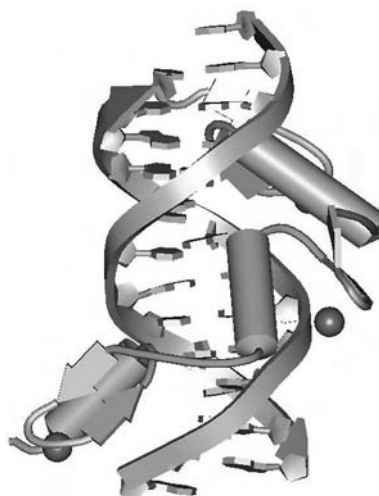


Figure 7.9 The structure of the Zif268 transcription factor-DNA complex (Pavletich and Pabo, 1991). The zinc atoms in the three zinc fingers are shown as spheres.

basic lysine side-chains and guanine/cytosine bases that exactly specifies the highly conserved sequence CCG at this point along the DNA. Nuclear receptors, such as those for steroid hormones and retinoic acid, also commonly use zinc-containing motifs for recognition of their DNA response elements, and a number of NMR and crystal structures of such complexes have been reported. Almost all of the response elements contain the half-site consensus sequences d(AGGTCA) or d(AGAACA), with the receptors binding as homo- or heterodimers. The zinc motif in them consists of a pair of α -helices linked by zinc coordination through C-terminal loops (Fig. 7.11). One helix binds in the DNA major groove at each half-site (Fig. 7.12), while the role of the second helix is to maintain the overall structure (Luisi *et al.*, 1991; Schwabe *et al.*, 1993; Zhao *et al.*, 2000). The side-chains from the recognition helix participate in normal direct readout interactions; the DNA itself retains B-DNA form, though typically local distortions to roll and propeller twist have been observed.

Zinc fingers have been used in the design of synthetic DNA recognition molecules with sequence specificity that can be altered at will (Choo and Isalan, 2000; Papworth, Kolasinska, and Minczuk, 2006). This approach complements that of recognition of the minor groove by dimeric polyamides (see Chap. 5), and has the potential advantage of greater synthetic accessibility, and with potential applicability to all possible target sites with equal facility. They have the disadvantage of being protein-like in size, and therefore may have problems in being taken up into cells.

A peptide containing three zinc fingers, found by screening a library of peptides, has been shown to bind to a nine base-pair sequence in the BCR-ABL oncogene (Choo, Sánchez-García, and Klug, 1994), with an equilibrium constant of 6×10^{-7} M. This binding affinity is at least an order of magnitude higher than to other control sequences. The original strategy did not effectively recognize all possible bases within a triplet, with the 5' one being optimally adenine or cytosine, possibly as a result of sequence context effects. This problem was overcome (Isalan, Choo, and Klug, 1998;



Figure 7.10 The structure of the GAL4 transcription factor-DNA complex (Marmorstein *et al.*, 1992).

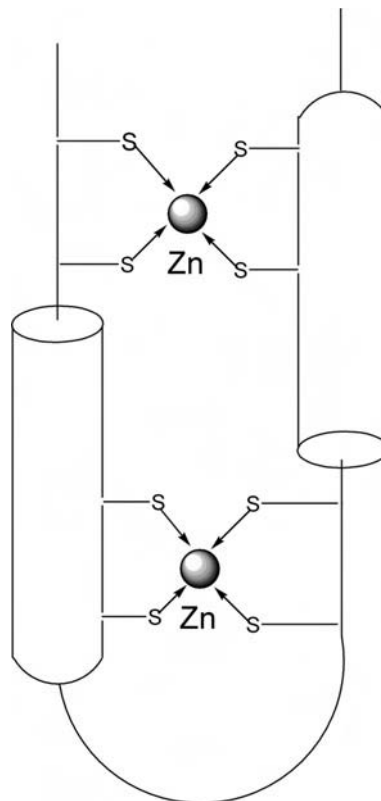


Figure 7.11 The arrangement for two zinc fingers as found in the nuclear receptor family of transcription factors. The fingers are formed between helices and loops, and the zinc atoms are coordinated by cysteine residues.



Figure 7.12 The structure of the estrogen receptor-DNA complex (Schwabe *et al.*, 1993), with the zinc atoms in the two zinc fingers shown as shaded spheres.

Isala, Klug, and Choo, 2001) by detailed considerations of several zinc finger crystal structures that show synergistic recognition from adjacent zinc fingers. As a result, it is possible to start devising a general recognition code for zinc-finger-directed DNA sequence specificity. Addition of a dimerization motif can enable more than three zinc fingers to be assembled together without loss of affinity. For example, a four-finger protein has been successfully constructed (Wolfe, Ramme, and Pabo, 2000) using the GCN4 leucine zipper dimerization motif (see below), which recognizes 10 base pairs.

The use of zinc finger motifs to specifically recognize any desired DNA sequence is increasingly being applied to a wide range of biological problems (Papworth, Kolasinska, and Minczuk, 2006), and several websites are now offering facilities for zinc finger design for this purpose (such as that of the Zinc Finger Consortium at <http://www.zincfingers.org/>).

7.5 Other Major Groove Recognition Motifs

A number of eukaryotic transcription factors contain the basic “leucine-zipper” recognition and dimerization element, as found in the yeast transcription factor GCN4 (Ellenberger *et al.*, 1992), which consists of a very long continuous, almost straight α -helix with regularly repeating leucine residues (Fig. 7.13). Two such helices interact together to form a parallel coiled-coil. There are a large number of interactions with the DNA backbone from the basic side-chains, as well as specific, direct readout ones. The orientation of the α -helix within the DNA major groove is critical for these contacts to take place, with the helix being approximately perpendicular to the phosphodiester backbone.

A distinct DNA-binding motif is used by the transcription activator E2 from papilloma virus (Hegde *et al.*, 1992), with a dimeric antiparallel β -barrel which delivers a pair of α -helices to the major groove. Protein-DNA contacts involve direct side-chain interactions with the bases, as well as indirect backbone contacts. A distinguishing feature of the structure is the smooth bending of the DNA around the β -barrel, with a radius of curvature of 45 Å. This is less extreme than the 90° bend found (Schultz, Shields, and Steitz, 1991) in the DNA of the complex with the dimeric HTH domain *E. Coli* transcription activator protein (CAP). The bending in the CAP complex, discussed further below, is due to a large number of interactions with DNA phosphate groups, which serve to position the recognition helix correctly in the major groove (Fig. 7.14). A totally distinct nonhelix recognition motif has been found in the structures of some bacterial repressor proteins, typified by the *met* J repressor/operator complex (Somers and Phillips, 1992), where double-stranded antiparallel β -ribbons are the major-groove recognition elements. Side-chains from the ribbons interact directly with the DNA phosphates and bases; the side-chain–base direct readout mechanism remains the same as that observed with the other repressor/operator complexes outlined above. Other analogous types of folds have been found in the structures of several DNA repair enzymes.

7.6 Minor-Groove Recognition

7.6.1 Recognition of B-DNA

Sequence-specific proteins generally do exploit the major groove to a greater extent than the minor one. This is not always solely on account of its greater information potential. The ability of an α -helix to snugly fit into the major groove is also a factor. This led to a view in the past that the minor groove is of little importance for

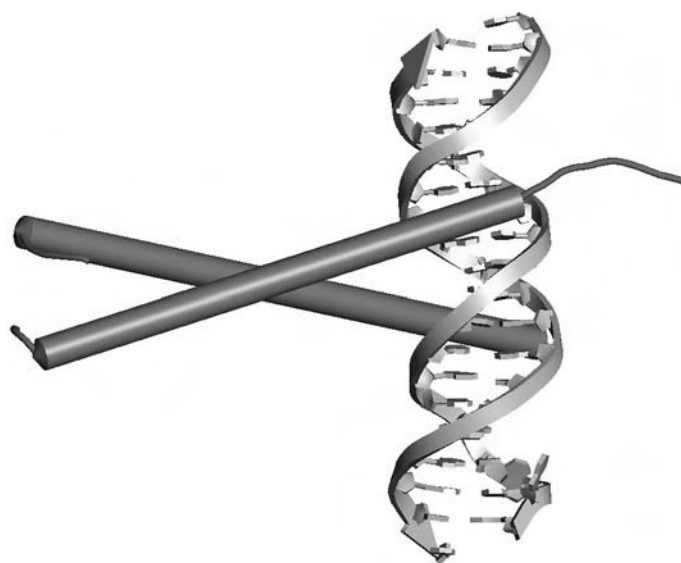


Figure 7.13 The structure of the yeast transcription factor GCN4-DNA complex (Ellenberger *et al.*, 1992).

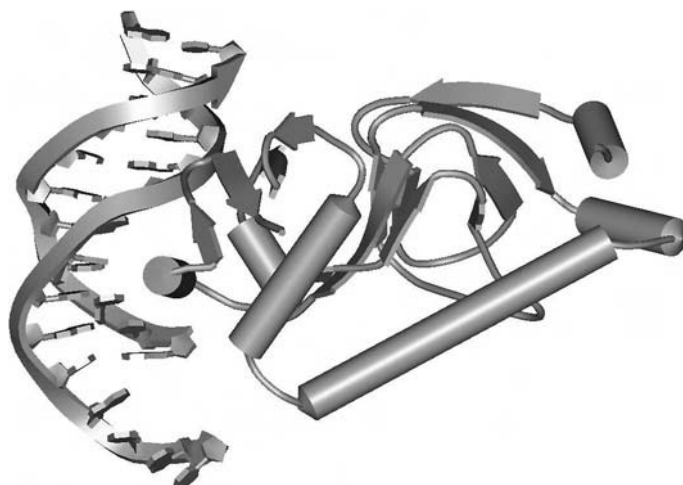


Figure 7.14 The structure of the *E. Coli* transcription activator protein (CAP)-DNA complex, showing one CAP unit with its recognition helix in the major groove of the bent DNA (Schultz, Shields, and Steitz, 1991).

Table 7.4 Crystal Structures of Some Protein–DNA Complexes with Miscellaneous Binding Motifs

Protein	PDB ID code	NDB ID code
Met J	1MJM	PD0245
E2	2BOP	PDV001
P53	1TSR	PDR022
Ku	1JEY	PD0220
NF- κ B	2RAM	PDR051
Nucleosome, 2.8 Å resolution	1AOI	PD0001
Nucleosome, 1.94 Å resolution	1KX5	PD0287
Tetranucleosome	1ZBB	PD0639

protein recognition. However there are now a large number of examples where interactions of side-chains in the minor groove are significant components of the overall protein-DNA stabilization (Table 7.5). An arginine side-chain of the 434 repressor (Aggarwal *et al.*, 1988) bridges to bases and phosphate groups via water molecules. Homeodomain structures (see, e.g., Kissinger *et al.*, 1990; Wolberger *et al.*, 1991; Li *et al.*, 1995; Liet *al.*, 1998; Tan and Richmond, 1998) have extended N-terminal arms which lie in the DNA minor groove, making extensive base and backbone contacts to the A/T regions of the operator sequence. The hydrogen-bonding interactions that the arginine side-chains in these complexes make with the O2 atom of thymines, are closely analogous to the mode of interaction shown by minor-groove binding drugs (Morávek, Neidle, and Schneider, 2002). These interactions, together with the indirect readout of the dimensions of the groove itself by van der Waals interactions involving the side-chains, is a significant factor in the preference shown by homeodomains for A/T-rich sites on DNA, again analogous to the preferences shown by the drugs. These N-terminal sequences are also related to the proposal (Suzuki, 1989) that

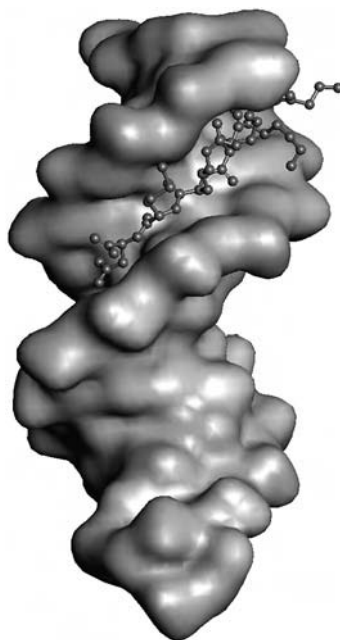
Table 7.5 Structures of Selected Protein–DNA Complexes Involving Minor-Groove Recognition

Protein	PDB ID code	NDB ID code
Hin recombinase	1HCR	PDE009
TBP	1D3U, 1QN4	PD0070, PD0154
TBP-TFIIB complex	1VOL	PDT032
HMG-D	1QRV	PD0110

the repeating sequence SPKK (serine-proline-lysine/arginine-lysine/arginine), which occurs in, for example, the N-terminus of some histone proteins, also binds in the minor groove at A/T-rich regions.

The crystal structure of the HTH-motif Hin recombinase enzyme bound to a 14 base-pair A/T-rich DNA sequence (Feng, Johnson, and Dickerson, 1994) shows that the straight B-DNA helix in this structure has a narrow minor groove in its A/T region, in which the N-terminus of the enzyme resides (Fig. 7.15). There are several amino acid–base edge specific contacts, such as an arginine to N3 of an adenine. More unusually, the main chain amide of this arginine is in hydrogen-bonding contact with O2 of a thymine. The pattern of hydrogen-bonding in this A/T-rich region is reminiscent of netropsin binding (see Chap. 5). The C-terminus of this enzyme also lies in the minor groove, entering at the other end of the DNA duplex.

Loops can also bind in the minor groove, as found in the solution NMR structure of the Mu repressor protein-DNA complex (Wojciak, Iwahara, and Clubb, 2001), which has a HTH motif containing an additional “wing” loop between helix 2 and 3. The wing is flexible in the absence of DNA, but in the complex is found inserted in

**Figure 7.15** A view of part of the structure of the hin recombinase-DNA complex, showing the N-terminus residing in the minor-groove surface of the DNA (Feng, Johnson, and Dickerson, 1994).

the minor groove, where two lysine side-chains make extensive contacts with, in particular, adenine and thymine base-edge atoms.

7.6.2 The Opening-up of the Minor Groove by TBP

The general transcription factor complex TFIID plays a key role in the initiation of transcription in eukaryotic cells. It functions by binding a component protein, TBP, to the “TATA box” sequence upstream of the start of transcription. This has been shown by chemical protection studies to bind in the minor groove at the 5′-TATA site itself, as well as in the major groove on the 3′ side of this site. There is a strong preference for 5′-TATA as compared to other A/T-containing sequences. The crystal structure of the TBP protein (Nikolov *et al.*, 1992), in the absence of DNA, shows a novel DNA-binding fold, with a symmetric α/β arrangement. It was initially considered that that the saddle-shaped arrangement of the α/β structure, with an extended concave surface, is ideally complementary to a B-form DNA duplex, and would be effectively wrapped around it. Mutagenesis studies have identified the key protein residues involved in DNA binding. They are arranged on the concave surface of the protein, as suggested by this interaction model.

However the actual structures of the TBP-DNA complexes show a very different arrangement for the DNA compared to this initial model (Kim, Nikolov, and Burley, 1993; Kim *et al.*, 1993). The DNA is dramatically bent, by $\sim 80^\circ$ so as to follow the concave curvature of the β -sheet, i.e., it is at right angles to the earlier model (Fig. 7.16). This structure has been observed in TATA-containing DNA sequences complexed with TBP from a wide range of organisms, so is clearly a general feature of TATA box recognition (Patikoglou *et al.*, 1999). Eight base pairs are in contact with the β -sheet saddle, and are in a non-B-DNA conformation. By contrast, analysis of complexes with longer DNA sequences shows that the DNA flanking the central site is in a B-like form, on both 3′ and 5′ sides. Structure determination at high-resolution

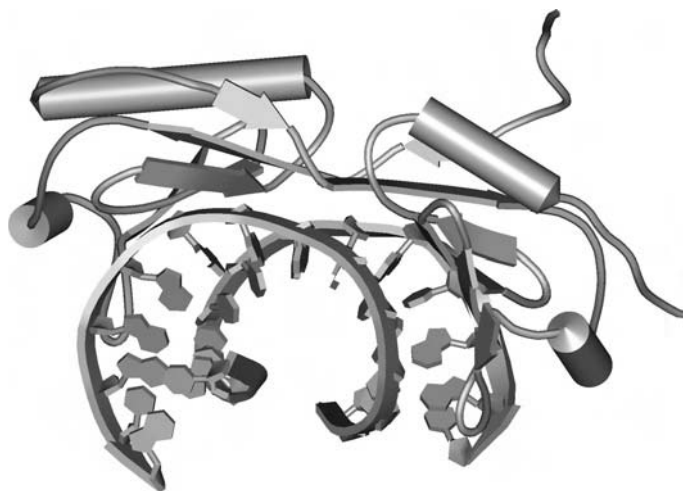


Figure 7.16 The structure of the TBP-DNA complex (Kim and Burley, 1994). Note the high degree of curvature of the DNA duplex.

(1.9 Å) of a TBP complex with the sequence d(TATAAAAG) has enabled a detailed view of the bent DNA structure to be obtained (Kim and Burley, 1994). The DNA is unwound by 105° over the seven base pairs, and has a greatly enlarged and flattened minor groove (with a maximum width of over 9 Å) to accommodate the β-sheet saddle. The major groove is highly compressed. There are large positive rolls, of up to 40° at the TA step in the 5'-TATA sequence, together with large propeller twists, of up to -39° for base pairs in the 5'-AAAA sequence. These large twists result in a pattern of A-tract major-groove bifurcated hydrogen bonds. Sugar puckers are in the C3'-*endo* family. The overall impression of the DNA conformation in the TATA box region is of distorted A-type form, with abrupt changes to B-DNA morphology immediately outside the box. The bending is achieved by pairs of phenylalanine residues that are inserted into the ends of the TATA box, and are stacked with several bases and base pairs. These bases become unstacked from the DNA helix and result in the observed kinking. There are few direct readout base-side-chain contacts, but numerous amino acid contacts with phosphate groups. Many of these, especially with lysine side-chains, are water-mediated.

Several ternary complexes of general transcription factors with the TBP-TATA complex have been reported. Those with TFIIA (Nikolov *et al.*, 1995) and TFIIB (Tan *et al.*, 1996) show that the TBP-TATA box structure seen in the binary complexes, is fully retained, strongly suggesting that these transcription factors bind to a preformed TATA box complex (Fig. 7.17), as does the negative cofactor NC2 in its TBP-DNA complex (Kamada *et al.*, 2001).

7.6.3 Other Proteins that Induce Bending of DNA

The ability of TBP to bend DNA on binding to its recognition sequence is shared by a number of other minor-groove regulatory proteins notably those containing the so-called HMG (high mobility group) box sequence-neutral DNA-binding domains.

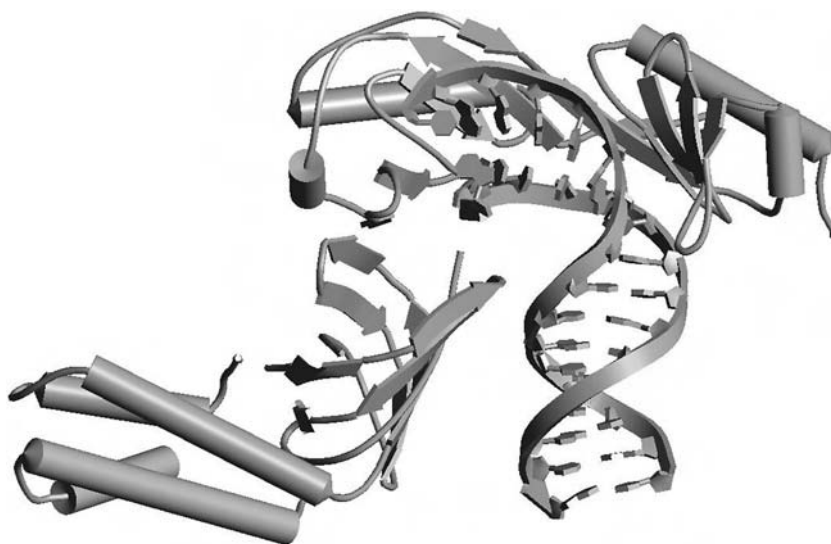


Figure 7.17 Structure of the TFIIA-TBP-DNA complex (Tan *et al.*, 1996). Note that the DNA structure is seen to revert to B-form outside the environment of the TBP saddle structure.

Such proteins include SRY, which determines the expression of human male-specific genes. NMR studies of HMG boxes (Weir *et al.*, 1993; Werner *et al.*, 1995; Allain *et al.*, 1999) show that their structure consists of three α -helices arranged in an L shape, with conserved basic arginines and lysines on the inner face of the L. The helices do not correspond to those in any (major-groove-binding) HTH arrangement, and thus constitute a novel DNA-binding motif. A crystal structure of a complex between HMG-D and an A/T-rich duplex decamer (Murphy, Sweet, and Churchill, 1999) shows that the bound DNA is very distorted from canonical B-DNA, and smoothly bent (Fig. 7.18). The protein binds to the minor groove, which is underwound, widened, and flattened so that the overall shape resembles A-DNA. This is reminiscent of the distortions induced by the TBP protein, and indeed there are significant similarities (Fig. 7.19), although in the DNA-HMG-D structure the minor



Figure 7.18 Structure of the HMG-D-DNA complex (Murphy, Sweet, and Churchill, 1999).

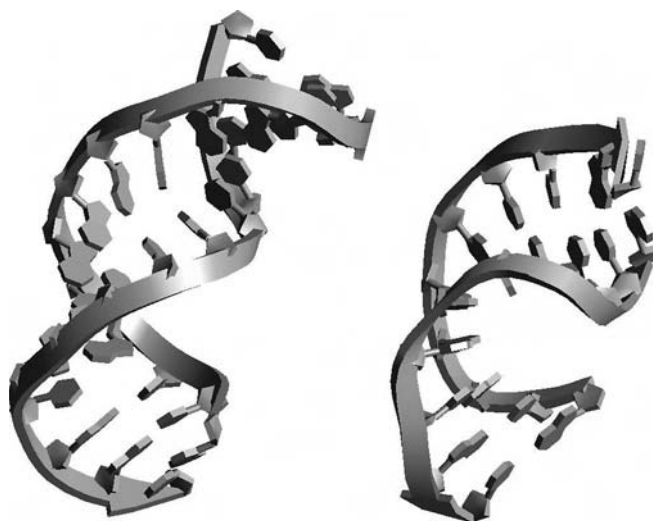


Figure 7.19 A comparison of the bent DNA duplexes in the TBP-TATA complex (Kim and Burley, 1994) and in the HMG-D-DNA complex (Murphy, Sweet, and Churchill, 1999).

groove is even wider, 12.4 Å. The crystal structures of the DNA complexes with the bacterial proteins HU (histone-like) and IHF (integration host factor), show that they are also analogous in inducing very large bends in the DNA even though the proteins are structurally dissimilar (Swinger and Rice, 2004, 2007) (Table 7.5).

The organization of almost two meters of DNA in the eukaryotic chromosome requires that it is packaged-up in a very tight manner (Luger and Richmond, 1998). The basic unit of packaging is the nucleosome core particle, which consists of an octamer of pairs of the four histone proteins H2A, H2B, H3, and H4, together with between 160 and 240 DNA base pairs, wound around the core. The nucleosome particle itself is the repeating unit for the further level of supercoiled organization in chromosomes. The crystal structure of the nucleosome core particle was initially determined at 7 Å resolution (Richmond *et al.*, 1984), and subsequently at 2.8 Å (Luger *et al.*, 1997), which enabled details such as DNA-bending, to be described. In this structure, there are 146 base pairs, which are wound around the octameric histone core, forming a flattened disklike superhelix (Fig. 7.20), such that

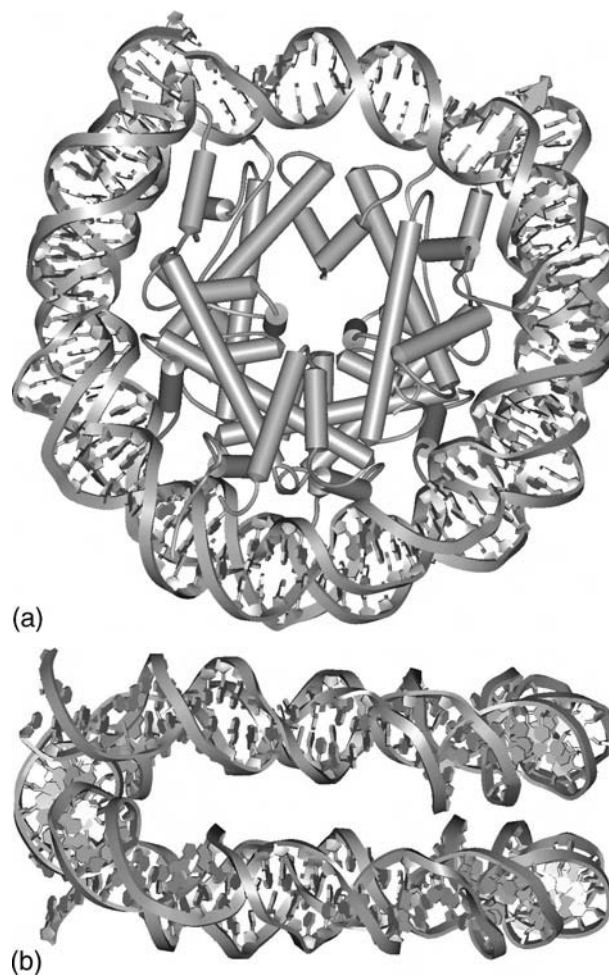


Figure 7.20 Two views of the nucleosome structure (Luger *et al.*, 1997), showing (a) the histone octamer surrounded by supercoiled DNA. (b) A view at right angles highlighting the disklike nature of the structure, which has almost two complete turns of bound supercoiled DNA.

the structure overall has pseudo twofold symmetry. There are 10.2 base pairs per turn and at this resolution the bending was observed to be altogether regular, with the DNA retaining B-like features. The smooth bending was seen at this resolution to be accomplished by relatively small but numerous deformations from ideality, in a sequence-dependent manner, analogous to changes observed in crystal structures of some short yet bent oligonucleotides. The more recent availability of highly diffracting crystals has enabled a remarkably high-resolution (1.9 Å) refinement to be undertaken (Richmond and Davey, 2003). This is comparable to the resolution reported for many native oligonucleotides, and thus has enabled an altogether more detailed comparative picture of local DNA deformations to be obtained than was possible at lower resolution. This shows that nucleosomal DNA is B-form overall, but bending is not uniform, due to (a) the effects of the histone proteins at various points, (b) anisotropic flexibility of the DNA, and (c) local sequence-dependent structural features, some of which have not been observed in native oligonucleotide crystal (or NMR) structures. This irregularity is apparent in Fig. 7.21, where we see an almost straight DNA tract followed by a more bent tract. The base-step morphology parameters such as roll, helical twist, and tilt vary in a sinusoidal manner along the 147 nucleotide length, with surprisingly large variations at different points (Fig. 7.22), which cumulatively result in double the curvature for the DNA than is actually required. The origin of this extra bending appears to be an accumulation of



Figure 7.21 A short stretch of the DNA, taken from the high-resolution crystal structure of the nucleosome (Richmond and Davey, 2003), showing the nonuniform nature of the DNA curvature in this region.

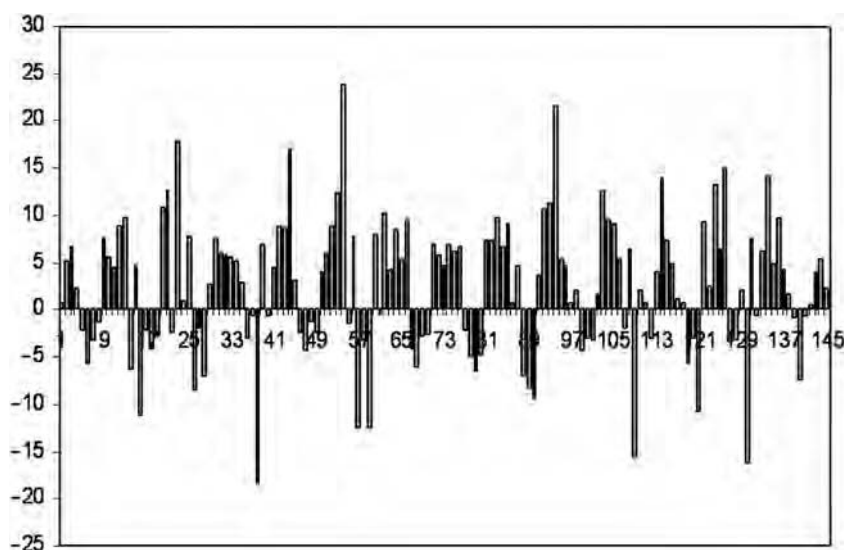


Figure 7.22 A plot of the base-step roll angles for the DNA in the high-resolution nucleosome crystal structure (Richmond and Davey, 2003). The progression of base steps is indicated along the *x*-axis and roll angles are shown on the *y*-axis.

excessive roll. There is a small preference for bending in the direction of the major groove, characterized by smooth DNA deformations; by contrast the minor-groove bending has extensive kinking. This sequence-dependent effect occurs at CA base steps, and is associated with significant deviations from B-form DNA mean values for a number of parameters describing base-step and base-pair morphology. The crystal structure of a human DNA sequence in a nucleosome core structure (Tsunaka *et al.*, 2005) shows a number of differences in the DNA path as it winds around the histone core, compared to the structure described above, although the overall arrangement is closely similar to it (and to all other nucleosome structures, studied from a range of eukaryotic organisms). These differences may be a consequence of DNA–DNA packing interactions in the crystals of the human form.

The next level of organization in eukaryotic chromosomes is the chromatin fiber. The crystal structure, at low (9 Å) resolution (Schalch *et al.*, 2005) of the tetranucleosome has revealed detail about how individual nucleosomes may be organized in the fiber. Each nucleosome is connected by 20 base-pair linkers, in a supercoiled zigzag-like arrangement (Fig. 7.23). The overall structure has been used to build plausible models for the chromatin fiber having a continuous array of nucleosome units.

7.7 DNA-Bending and Protein Recognition

X-ray crystallography, NMR and other techniques have shown that DNA structures in the absence of bound drug or protein can have a degree of flexibility around the mean. This is reflected in sequence-dependent backbone, sugar pucker, base-step, and base-pair features, although the classic features of canonical forms are often observed in both native and protein-bound oligonucleotide structures.

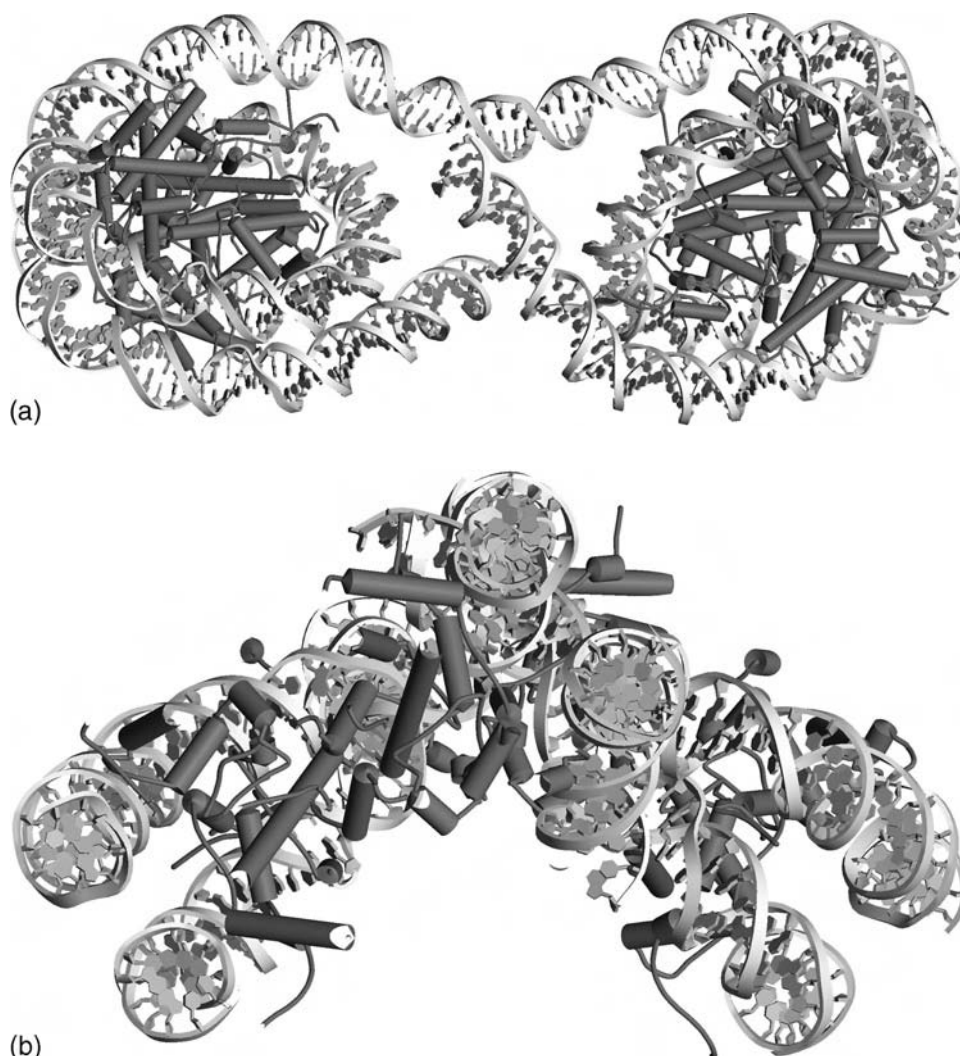


Figure 7.23 Two views of the crystal structure of the tetranucleosome (Schalch *et al.*, 2005) (a) looking onto the histone cores, and (b) showing an enlarged view of the angular arrangement between individual units.

There is considerably greater diversity of DNA structure in protein complexes than in native DNA structures, largely because the binding energy for a protein–DNA interaction is sufficient to readily overcome the barriers to local DNA deformability. Hence these can lead to large-scale changes in DNA structure such as bending, so as to optimize interactions, should such changes be needed. Many proteins, notably HTH and zinc fingers, do not normally need to distort DNA in order to produce the required protein contacts – some notable exceptions are described below. Sometimes though when distortions do occur they can blur the distinctions between B- and A-form DNA structural types so that it is no longer possible to unequivocally assign the resulting DNA structure to one or the other form. This is especially apparent when large changes in parameters such as groove width, helical twist, and base-pair inclination to the helix axis are observed. For example, the nature of the structure

of DNA in the Zif268-DNA zinc-finger complex has been controversial, with both A- and B-type helices being suggested. The crystal structure of the complex (Pavletich and Pabo, 1991) shows that the DNA adopts an overall B-type form, although there are a number of local features that are intermediate between A and B. In such a case, the simple canonical DNA classifications are by themselves of limited use and are best accompanied by more detailed information describing particular features.

The DNA in the *met* J complex (Somers and Phillips, 1992) has a straight 10 base-pair central segment, with bends at each end as a result of groove width changes necessary for both effective protein dimer formation and protein-DNA interaction. In the structure of the *trp* repressor/operator complex (Otwinowski *et al.*, 1988; Haran, Joachimiak, and Sigler, 1992) the DNA is appreciably distorted from canonical B-form as a result of numerous small, cumulative local changes to, in particular roll and slide. These together with changes in groove width and backbone geometry from canonical B-DNA are necessary in order to achieve the large number of indirect read-out contacts that produce specificity. The DNA is bent and its surface closely follows the contours of the protein surface. Rather greater DNA bending is required in the papillomavirus-1 E2 DNA complex (Hegde *et al.*, 1992), where there is smooth DNA bending over the β -barrel recognition motif, in order to expose the major groove to protein side-chains. Groove widths are compressed on the side of the DNA that faces protein, with alterations in propeller twist and roll angles contributing to these width changes. The prokaryotic transcription activator CAP, which acts as a dimer, has a consensus DNA binding site 22 base-pairs long with the essential bases toward the end of the sequence. This is far too long a DNA sequence to be contacted by the relatively small CAP dimer on the basis of the DNA remaining linear. The structures of the CAP-DNA complexes show (Schultz, Shields, and Steitz, 1991; Benoff *et al.*, 2002) that the DNA has solved this problem by bending by $\sim 90^\circ$ in order to effectively wrap around the protein dimer. This bending, or kinking, is produced almost entirely at two base-pair steps, one on each side of the twofold axis of the complex, by roll angles of $\sim 40^\circ$ and a twist angle of $\sim 20^\circ$ (Napoli *et al.*, 2006). It is likely that this bending is an important aspect of transcriptional activation by CAP and other analogous proteins. There is a very close structural correspondence between CAP and other architectural proteins (Werner and Burley, 1997) such as TBP and the globular domain of histone H5 (Ramakrishnan *et al.*, 1993) (which is involved in higher-level nucleosome organization). These suggest that the ability of this particular type of HTH protein to bend DNA sequences is of wide structural and functional significance.

In the light of the observations of DNA deformability in these and other protein complexes, the question arises as to their relevance to the numerous studies on native DNA structures. The answer is not clear-cut, and depends on the particular structures and sequences involved. It is the case that native DNA structures can demonstrate inherent sequence-dependency which is also found in their protein complexes, provided the changes in structure are not profound. Examples are provided by, on the one hand, the *trp* operator site (Shakked *et al.*, 1994), which shares common features of structure and hydration in the native DNA structure and in this protein-DNA complex, and on the other hand, by the human papillomavirus type 6 E2 protein, which shows considerable bending in its DNA complex (Hooley *et al.*, 2006), although the A-tract-containing DNA sequence does show a predisposition to bending in its native structure. Larger-scale DNA conformational differences such as are seen in the TBP and other minor-groove binding protein complexes, can still be related to native DNA structures since these are often describable in terms of simple concerted changes from and to standard forms of

DNA. Thus the structure of the DNA in the TATA box may be thought of as a canonical A-form with just a change in glycosidic angle (Guzikevich-Guerstein and Shakked, 1996). It is striking that this and other small variants of the A-form (Lu, Shakked, and Olson, 2000), when embedded within B-DNA, can produce bent and curved structures suitable for binding to a wide range of proteins (and small molecules such as cis-platinum – see Chap. 5 and Sect. 7.8). Such bending was foreseen over 25 years ago, as a natural consequence of junctions between canonical A- and B-forms (Selsing *et al.*, 1979). It was less obvious until comparatively recently, that, *both* forms would play key roles in the functioning of DNA.

7.8 Protein-DNA-Small Molecule Recognition

Chapter 5 has detailed the many of the binary structures between small molecules (often drugs), and DNA. These structures have often provided insights into biological activity and structure–activity relationships, but only rarely have structures been determined that enable a fuller picture of function to be obtained. This is the case for the sequence-specific polyamides, which are still the only class of DNA-binding small molecule for which there is structural information relevant to chromatin. The crystal structures of three pyrrole-imidazole polyamide ligand–nucleosome complexes have been determined (Suto *et al.*, 2003), one at a resolution of 2.3 Å. These polyamides form antiparallel hairpin structures, recognizing DNA sequence via the minor groove, as has been observed in their binary DNA complexes. The same arrangement is seen in the nucleosome complexes (Fig. 7.24), with a number of polyamide molecules found to be bound to each nucleosome on the exterior face, away from the histone octamer core. The observation of bound ligand is itself significant, demonstrating that such molecules can indeed bind to nucleosomal DNA when it is associated with histones. A number of distortions to the native nucleosome structure are apparent, but only to the DNA and not to the histone core. These changes are largely increases in minor-groove width, which at some points are local and at others are transmitted along the sequence, so that some DNA-histone octamer contacts become loosened. One significant effect of polyamide binding is to block temperature-induced histone repositioning – this is likely to have consequences for chromatin function, not least remodeling during transcription. Polyamides have also been engineered to target the gap between the two DNA coils in a nucleosome, the “supergroove” (Edayathumangalam *et al.*, 2004). Crystallography has shown that this hairpin dimer (with two polyamide dimers linked by a short polyethylene glycol chain), does bind in the predicted manner, in accord with the finding that this ligand hinders nucleosome dissociation.

There are few other drug–DNA–protein ternary complexes for which there is structural information. The HMG domain complexed with DNA bound to the anticancer drug cis-platinum (see Chap. 5) shows a highly bent DNA (Fig. 7.25), with the minor

Table 7.6 Crystal Structures of Selected Protein–DNA–Drug Complexes

Protein-DNA	Small molecule/lesion	PDB ID code	NDB ID code
Nucleosome	ImPyPyPy- γ -PyPyPyPy- β -Dp	1M19	PD0329
HMG domain + DNA	Cis-platinum	1CKT	PD0051
Human topoisomerase I	Topotecan	1K4T	PD0256

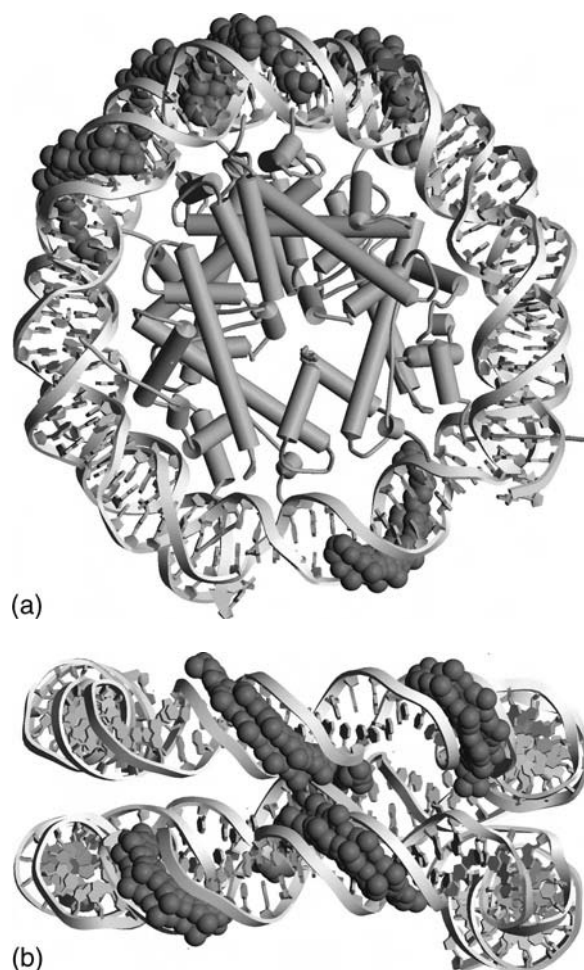


Figure 7.24 Two views from the crystal structure (Suto *et al.*, 2003) of one of the three nucleosome-polyamide complexes (PDB code 1M19), with polyamide molecules shown in space-filling representation. The distinctive dimeric side-by-side arrangement of the ligand at each binding site is clearly visible.

groove resembling A-DNA (Ohndorf *et al.*, 1999). The DNA topoisomerase enzymes are responsible in both prokaryotes and eukaryotes for changing the supercoiling of DNA during replication or transcription. They do this by first cleaving one or both DNA strands, then producing the desired topological change in DNA, and finally resealing the breaks. A variety of drugs, ranging from antibacterial agents to anticancer drugs, can interfere with topoisomerase function, often by preventing the relegation step from occurring. The crystal structure of human topoisomerase I in a ternary complex with the anticancer drug topotecan bound to a 22-mer DNA (Staker *et al.*, 2002), shows the drug bound in an intercalation-like pocket (Figs. 7.26 and 7.27) of the straight B-form DNA helix. It thus acts as a conventional intercalating molecule in moving one part of the DNA sequence relative to the other, down by one base pair, and thus interfering with normal topoisomerase-substrate function. Crystal structures of ternary complexes involving the related drug camptothecin with topoisomerase I



Figure 7.25 A view of the crystal structure of the HMG-DNA-cis-platinum ternary complex (Ohndorf *et al.*, 1999). The platinum drug is shown highlighted in space-filling representation.

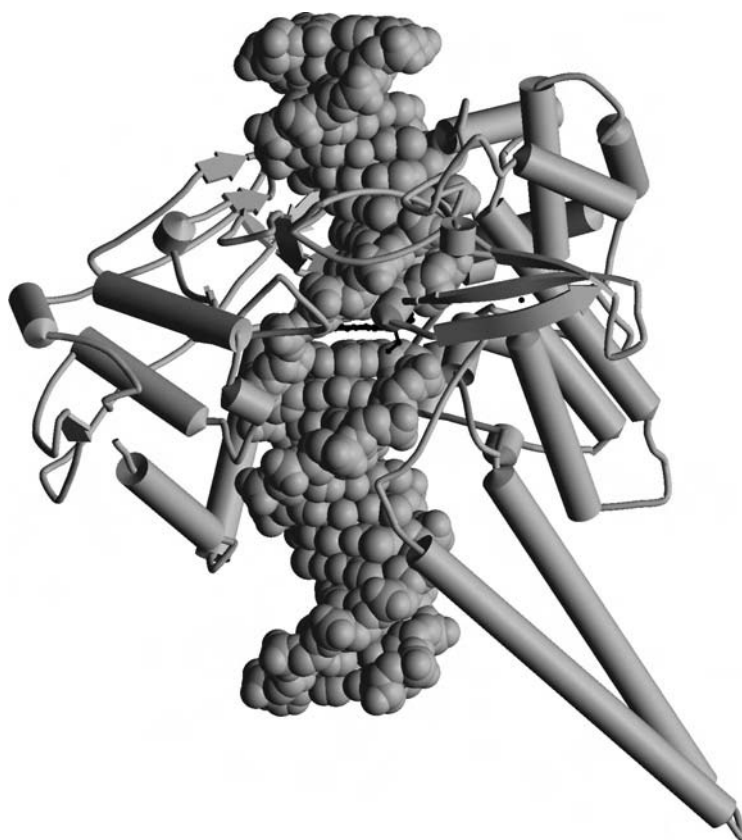


Figure 7.26 A view of the crystal structure of the ternary complex involving the drug topotecan bound to a 22-mer DNA and DNA topoisomerase I (Staker *et al.*, 2002). The drug molecule is in the gap in the middle of the DNA, and is drawn in black in ball-and-stick mode.

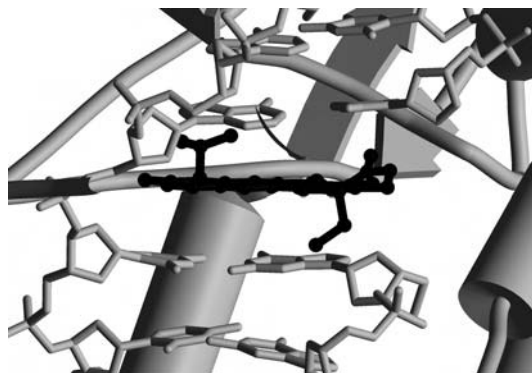


Figure 7.27 A detailed view of the binding site in the ternary topoisomerase complex (Staker *et al.*, 2002), with the topotecan molecule shown shaded in black. The break in the DNA strand on the right-hand side is clearly visible.

mutants that confer resistance to the drug (Chrencik *et al.*, 2004) have revealed, for example, changes in drug-enzyme contacts, which have been able to rationalize the resistance observations. Further crystal-structure analyses of topoisomerase I ternary complexes with a structurally diverse range of topoisomerase inhibitors (Staker *et al.*, 2005), have provided data for the rational design of novel types of agent of this general class.

Useful Websites

http://www.biochem.ucl.ac.uk/bsm/prot_dna/prot_dna_cover.html

This provides summaries of the major protein-DNA families, and the relevant features in many individual structures.

<http://www.biochem.ucl.ac.uk/bsm/DNA/server/>

The server provides tools for the analysis of the protein interface of any desired protein–nucleic acid complex.

<http://gibk26.bse.kyutech.ac.jp/jouhou/3dinsight/recognition.html>

The Protein-DNA Recognition Database, provides structural databases and prediction tools using, for example, sequence data to predict if a protein is DNA-binding or not.

References

- Aggarwal, A. K., Rodgers, D. W., Drottar, M., Ptashne, M., and Harrison, S. C. (1988). Recognition of a DNA operator by the repressor of phage 434: a view at high resolution. *Science*, **242**, 899–907.
- Aishima, J., Gitti, R. K., Joan, J. E., Gan, H. H., Schlick, T., and Wolberger, C. (2002). A Hoogsteen base pair embedded in undistorted B-DNA. *Nucleic Acids Res.*, **23**, 5244–5252.
- Allain, F. H.-T., Yen, Y.-M., Masse, J. E., Schultze, P., Dieckmann, T., Johnson, R. C., and Feigon, J. (1999). Solution structure of the HMG protein NHP6A and its interaction with DNA reveals the structural determinants for non-sequence-specific binding. *EMBO J.*, **18**, 2563–2579.
- Anderson, W. F., Ohlendorf, D. H., Takeda, Y., and Matthews, B. W. (1981). Structure of the cro repressor from bacteriophage lambda and its interaction with DNA. *Nature*, **290**, 754–758.
- Beamer, L. J. and Pabo, C. O. (1992). Refined 1.8 Å crystal structure of the lambda repressor-operator complex. *J. Mol. Biol.*, **227**, 177–196.

- Benoff, B., Yang, H., Lawson, C. L., Parkinson, G., Liu, J., Blatter, E., Ebright, Y. W., Berman, H. M., and Ebright, R. H. (2002). Structural basis of transcription activation: the CAP- α CTD-DNA complex. *Science*, **297**, 1562–1566.
- Billeter, M., Güntert, P., Luginbühl, P., and Wüthrich, K. (1996). Hydration and DNA recognition by homeodomains. *Cell*, **85**, 1057–1065.
- Choo, Y. and Isalan, M. (2000). Advances in zinc finger engineering. *Curr. Opin. Struct. Biol.*, **10**, 411–416.
- Choo, Y., Sánchez-García, I., and Klug, A. (1994). In vivo repression by a site-specific DNA-binding protein designed against an oncogenic sequence. *Nature*, **372**, 642–645.
- Chrencik, J. E., Staker, B. L., Burgin, A. B., Pourquier, P., Pommier, Y., Stewart, L., and Redinbo, M. R. (2004). Mechanisms of camptothecin resistance by human topoisomerase I mutations. *J. Mol. Biol.*, **339**, 773–784.
- Edayathumangalam, R. S., Weyermann, P., Gottesfeld, J. M., Dervan, P. B., and Luger, K. (2004). Molecular recognition of the nucleosomal “super groove”. *Proc. Natl. Acad. Sci. USA*, **101**, 6844–6869.
- Ellenberger, T. E., Brandl, C. J., Struhl, K., and Harrison, S. C. (1992). The GCN4 basic region leucine zipper binds DNA as a dimer of uninterrupted α helices: crystal structure of the protein-DNA complex. *Cell*, **71**, 1223–1237.
- Feng, J. A., Johnson, R. C., and Dickerson, R. E. (1994). Hin recombinase bound to DNA: the origin of specificity in major and minor groove interactions. *Science*, **263**, 348–355.
- Fraenkel, E. and Pabo, C. O. (1998). Comparison of X-ray and NMR structures for the Antennapedia homeodomain-DNA complex. *Nature Struct. Biol.*, **5**, 692–697.
- Fraenkel, E., Rould, M. A., Chambers, K. A., and Pabo, C. O. (1998). Engrailed homeodomain-DNA complex at 2.2 Å resolution: a detailed view of the interface and comparison with other engrailed structures. *J. Mol. Biol.*, **284**, 351–361.
- Gehring, W. J., Qian, Y. Q., Billeter, M., Furukubo-Tokunaga, K., Schier, A., Resendez-Perez, D., Affolter, M., Otting, G., and Wüthrich, K. (1994). Homeodomain-DNA recognition. *Cell*, **78**, 211–223.
- Guzikevich-Guerstein, G., and Shakked, Z. (1996). A novel form of the DNA double helix imposed on the TATA-box by the TATA-binding protein. *Nature Struct. Biol.*, **3**, 32–37.
- Haran, T. E., Joachimik, A., and Sigler, P. B. (1992). The DNA target of the trp repressor. *EMBO J.*, **11**, 3021–3030.
- Harrison, S. C. (1991). A structural taxonomy of DNA-binding domains. *Nature*, **353**, 715–719.
- Hegde, R. S., Grossman, S. R., Laimonis, L. A., and Sigler, P. B. (1992). Crystal structure at 1.7 Å of the bovine papillomavirus-1 E2 DNA-binding domain bound to its DNA target. *Nature*, **359**, 505–512.
- Hooley, E., Fairweather, V., Clarke, A. R., Gaston, K., and Brady, R. L. (2006). The recognition of local DNA conformation by the human papillomavirus type 6 E2 protein. *Nucleic Acids Res.*, **34**, 3897–3908.
- Isalan, M., Choo, Y., and Klug, A. (1998). Comprehensive DNA recognition through concerted interactions from adjacent zinc fingers. *Biochemistry*, **37**, 12026–12033.
- Isalan, M., Klug, A., and Choo, Y. (2001). A rapid, generally applicable method to engineer zinc fingers illustrated by targeting the HIV-1 promoter. *Nature Biotech.*, **19**, 656–660.
- Kamada, K., Shu, F., Chen, H., Malik, S., Stelzer, G., Roeder, R. G., Meisterernst, M., and Burley, S. K. (2001). Crystal structure of negative cofactor 2 recognizing the TBP-DNA transcription complex. *Cell*, **106**, 71–81.
- Kim, J. L. and Burley, S. K. (1994). 1.9 Å resolution refined structure of TBP recognizing the minor groove of TATA-AAAG. *Nature Struct. Biol.*, **1**, 638–653.
- Kim, J. L., Nikolov, D. B., and Burley, S. K. (1993). Co-crystal structure of TBP recognizing the minor groove of a TATA element. *Nature*, **365**, 520–527.
- Kim, Y., Geiger, J. H., Hahn, S., and Sigler, P. B. (1993). Crystal structure of a yeast TBP/TATA-box complex. *Nature*, **365**, 512–520.
- Kissinger, C. R., Liu, B., Martin-Blanco, E., Kornberg, T. B., and Pabo, C. O. (1990). Crystal structure of an engrailed homeodomain-DNA complex at 2.8 Å resolution: a framework for understanding homeodomain-DNA interactions. *Cell*, **63**, 579–590.
- Klug, A. and Rhodes, D. (1987). Zinc fingers: a novel protein motif for nucleic acid recognition. *Trends Biochem. Sci.*, **12**, 464–469.
- Kummerfeld, S. K. and Teichmann, S. A. (2006). DBD: a transcription factor prediction database. *Nucleic Acids Res.*, **34**, D74–81.
- Labeets, L. A. and Weiss, M. A. (1997). Electrostatics and hydration at the homeodomain-DNA interface: chemical probes of an interfacial water cavity. *J. Mol. Biol.*, **269**, 113–128.
- Laity, J. H., Lee, B. M., and Wright, P. E. (2001). Zinc finger proteins: new insights into structural and functional diversity. *Curr. Opin. Struct. Biol.*, **11**, 39–46.
- Li, T., Jin, Y., Vershon, A. K., and Wolberger, C. (1998). Crystal structure of the MATA1/MATalpha2 homeodomain heterodimer in complex with DNA containing an A-tract. *Nucleic Acids Res.*, **26**, 5707–5718.

- Li, T., Stark, M. R., Johnson, A. D., and Wolberger, C. (1995). Crystal structure of the MATa1/MAT alpha 2 homeodomain heterodimer bound to DNA. *Science*, **270**, 262–269.
- Lu, X.-J., Shakked, Z., and Olson, W. K. (2000). A-form conformational motifs in ligand-bound DNA structures. *J. Mol. Biol.*, **300**, 819–840.
- Luger, K. and Richmond, T. J. (1998). DNA binding within the nucleosome core. *Curr. Opin. Struct. Biol.*, **8**, 33–40.
- Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F., and Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, **389**, 251–260.
- Luisi, B., Xu, W. X., Otwinowski, Z., Freedman, L. P., Yamamoto, K. R., and Sigler, P. B. (1991). Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA. *Nature*, **352**, 497–505.
- Luscombe, N. M., Austin, S. E., Berman, H. M., and Thornton, J. M. (2000). An overview of the structures of protein-DNA complexes. *Genome Biol.*, **1**, 1.
- Luscombe, N. M., Laskowski, R. A., and Thornton, J. M. (2001). Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res.*, **29**, 2860–2874.
- Mandel-Gutfreund, Y., Schueler, O., and Margalit, H. (1995). Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. *J. Mol. Biol.*, **253**, 370–382.
- Marmorstein, R., Carey, M., Prashne, M., and Harrison, S. C. (1992). DNA recognition by GAL4: structure of a protein-DNA complex. *Nature*, **356**, 408–414.
- Matthews, B. W. (1988). Protein-DNA interaction. No code for recognition. *Nature*, **335**, 294–295.
- Morávek, Z., Neidle, S., and Schneider, B. (2002). Protein and drug interactions in the minor groove of DNA. *Nucleic Acids Res.*, **30**, 1182–1191.
- Murphy, F. V., Sweet, R. M., and Churchill, M. E. A. (1999). The structure of a chromosomal high mobility group protein-DNA complex reveals sequence-neutral mechanisms important for non-sequence-specific DNA recognition. *EMBO J.*, **18**, 6610–6618.
- Nadassy, K., Wodak, S. J., and Janin, J. (1999). Structural features of protein-nucleic acid recognition sites. *Biochemistry*, **38**, 1999–2017.
- Nair, D. T., Johnson, R. E., Prakash, S., Prakash, L., and Aggarwal, A. K. (2004). Replication by human DNA polymerase- ϵ occurs by Hoogsteen base-pairing. *Nature*, **430**, 377–380.
- Nair, D. T., Johnson, R. E., Prakash, L., Prakash, S., and Aggarwal, A. K. (2006). An incoming nucleotide imposes an anti to syn conformational change on the templating purine in the human DNA polymerase- ϵ active site. *Structure*, **14**, 749–755.
- Napoli, A. A., Lawson, C. L., Ebright, R. H., and Berman, H. M. (2006). Indirect readout of DNA sequence at the primary-kink site in the CAP-DNA complex: recognition of pyrimidine-purine and purine-purine steps. *J. Mol. Biol.*, **357**, 173–183.
- Neri, D., Billeter, M., and Wüthrich, K. (1992). Determination of the nuclear magnetic resonance solution structure of the DNA-binding domain (residues 1 to 69) of the 434 repressor and comparison with the X-ray crystal structure. *J. Mol. Biol.*, **223**, 743–767.
- Nikolov, D. B., Chen, H., Halay, E. D., Usheva, A. A., Hisatake, K., Lee, D. K., Roeder, R. G., and Burley, S. K. (1995). Crystal structure of a TFIIB-TBP-TATA-element ternary complex. *Nature*, **377**, 119–128.
- Nikolov, D. B., Hu, S.-H., Lin, J., Gasch, A., Hoffmann, A., Horikoshi, M., Chua, N.-H., Roeder, R. G., and Burley, S. K. (1992). Crystal structure of TFIID TATA-box binding protein. *Nature*, **360**, 40–46.
- Ohndorf, U.-M., Rould, M. A., He, Q., Pabo, C. O., and Lippard, S. J. (1999). Basis for recognition of cisplatin-modified DNA by high-mobility-group proteins. *Nature*, **399**, 708–712.
- Otwinowski, Z., Schevitz, R. W., Zhang, R.-G., Lawson, C. L., Joachimiak, A., Marmorstein, R. Q., Luisi, B. F., and Sigler, P. B. (1988). Crystal structure of trp repressor/operator complex at atomic resolution. *Nature*, **335**, 321–329.
- Pabo, C. O. and Nekludova, L. (2000). Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? *J. Mol. Biol.*, **301**, 597–624.
- Passner, J. M., Ryoo, H. D., Shen, L., Mann, R. S., and Aggarwal, A. K. (1999). Structure of a DNA-bound ultrathin-thorax-extradenticle homeodomain complex. *Nature*, **397**, 714–719.
- Patikoglou, G. A., Kim, J. L., Sun, L., Yang, S.-H., Kodadek, T., and Burley, S. K. (1999). TATA element recognition by the TATA box-binding protein has been conserved throughout evolution. *Genes Develop.*, **13**, 3217–3230.
- Papworth, M., Kolasinska, P., and Minczuk, M. (2006). Designer zinc-finger proteins and their applications. *Gene*, **366**, 27–38.
- Pavletich, N. P. and Pabo, C. O. (1991). Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å. *Science*, **252**, 809–817.
- Pellegrini-Calace, M. and Thornton, J. M. (2005). Detecting DNA-binding helix-turn-helix structural motifs using sequence and structure information. *Nucleic Acids Res.*, **33**, 2129–2140.

- Piper, D. E., Batchelor, A. H., Chang, C.-P., Cleary, M. L., and Wolberger, C. (1999). Structure of a HoxB1-Pbx1 heterodimer bound to DNA: role of the hexapeptide and a fourth homeodomain helix in complex formation. *Cell*, **96**, 587–597.
- Ramakrishnan, V., Finch, J. T., Graziano, V., Lee, P. L., and Sweet, R. M. (1993). Crystal structure of globular domain of histone H5 and its implications for nucleosome binding. *Nature*, **362**, 219–223.
- Rastinejad, F., Perlmann, T., Evans, R. M., and Sigler, P. B. (1995). Structural determinants of nuclear receptor assembly on DNA direct repeats. *Nature*, **375**, 203–211.
- Richmond, T. J. and Davey, C. A. (2003). The structure of DNA in the nucleosome core. *Nature*, **423**, 145–149.
- Richmond, T. J., Finch, J. T., Rushton, B., Rhodes, D., and Klug, A. (1984). Structure of the nucleosome core particle at 7 Å resolution. *Nature*, **311**, 532–537.
- Sam, M. D., Cascio, D., Johnson, R. C., and Clubb, R. T. (2004). Crystal structure of the excisionase-DNA complex from bacteriophage lambda. *J. Mol. Biol.*, **338**, 229–240.
- Schalch, T., Duda, S., Sargent, D. F., and Richmond, T. J. (2005). X-ray structure of a tetranucleosome and its implications for the chromatin fibre. *Nature*, **436**, 138–141.
- Schultz, S. C., Shields, G. C., and Steitz, T. A. (1991). Crystal structure of a CAP-DNA complex: the DNA is bent by 90 degrees. *Science*, **253**, 1001–1007.
- Schwabe, J. W. R., Chapman, L., Finch, J. T., and Rhodes, D. (1993). The crystal structure of the estrogen receptor DNA-binding domain bound to DNA: how receptors discriminate between their response elements. *Cell*, **75**, 567–578.
- Seeman, N. C., Rosenberg, J. M., and Rich, A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl. Acad. Sci. USA*, **73**, 804–808.
- Selsing, E., Wells, R. D., Alden, C. J., and Arnott, S. (1979). Bent DNA: visualization of a base-paired and stacked A-B conformational junction. *J. Biol. Chem.*, **254**, 5417–5422.
- Shakked, Z., Guzikevich-Guerstein, G., Frolov, F., Rabinovich, D., Joachimiak, A., and Sigler, P. B. (1994). Determinants of repressor/operator recognition from the structure of the trp operator binding site. *Nature*, **368**, 469–473.
- Somers, W. S. and Phillips, S. E. V. (1992). Crystal structure of the met repressor-operator complex at 2.8 Å resolution reveals DNA recognition by beta-strands. *Nature*, **359**, 387–393.
- Staker, B. L., Feese, M. D., Cushman, M., Pommier, Y., Zembower, D., Stewart, L., and Burgin, A. G. (2005). Structures of three classes of anticancer agents bound to the human topoisomerase I-DNA covalent complex. *J. Med. Chem.*, **48**, 2336–2345.
- Staker, B. L., Hjerrild, K., Feese, M. D., Behnke, C. A., Burgin, A. B., and Stewart, L. (2002). The mechanism of topoisomerase I poisoning by a camptothecin analog. *Proc. Natl. Acad. Sci. USA*, **99**, 15387–15392.
- Suto, R. K., Edayathumangalam, R. S., White, C. L., Melander, C., Gottesfeld, J. M., Dervan, P. B., and Luger, K. (2003). Crystal structures of nucleosome core particles in complex with minor groove DNA-binding ligands. *J. Mol. Biol.*, **326**, 371–380.
- Suzuki, M. (1989). SPKK, a new nucleic acid-binding unit of protein found in histone. *EMBO J.*, **8**, 797–804.
- Swinger, D. K. and Rice, P. A. (2004). IHF and HU: flexible architects of bent DNA. *Curr. Opin. Struct. Biol.*, **14**, 28–35.
- Swinger, D. K. and Rice, P. A. (2007). Structure-based Analysis of HU-DNA Binding. *J. Mol. Biol.*, **365**, 1005–1016.
- Tan, S. and Richmond, T. J. (1998). Crystal structure of the yeast MATalpha2/MCM1/DNA ternary complex. *Nature*, **391**, 660–666.
- Tan, S., Hunziker, Y., Sargent, D. F., and Richmond, T. J. (1996). Crystal structure of a yeast TFIIA/TBP/DNA complex. *Nature*, **381**, 127–151.
- Tsunaka, T., Kajimure, N., Tate, S., and Morikawa, K. (2005). Alteration of the nucleosomal DNA path in the crystal structure of a human nucleosome core particle. *Nucleic Acids Res.*, **33**, 3424–3434.
- Weir, H. M., Kraulis, P. J., Hill, C. S., Raine, A. R. C., Laue, E. D., and Thomas, J. O. (1993). Structure of the HMG box motif in the B-domain of HMG1. *EMBO J.*, **12**, 1311–1319.
- Werner, M. H. and Burley, S. K. (1997). Architectural transcription factors: proteins that remodel DNA. *Cell*, **88**, 733–736.
- Werner, M. H., Huth, J. R., Gronenborn, A. M., and Clore, G. M. (1995). Molecular basis of human 46X,Y sex reversal revealed from the three-dimensional solution structure of the human SRY-DNA complex. *Cell*, **81**, 705–714.
- Wojciak, J. M., Iwahara, J., and Clubb, R. T. (2001). The Mu repressor-DNA complex contains an immobilized 'wing' within the minor groove. *Nature Struct. Biol.*, **8**, 84–90.
- Wolberger, C., Vershon, A. K., Liu, B., Johnson, A. D., and Pabo, C. O. (1991). Crystal structure of a MAT alpha 2 homeodomain-operator complex suggests a general model for homeodomain-DNA interactions. *Cell*, **67**, 517–528.

- Wolfe, S. A., Nekludova, L., and Pabo, C. O. (2000). DNA recognition by Cys2His2 zinc finger proteins. *Ann. Rev. Biophys.*, **29**, 183–212.
- Wolfe, S. A., Ramm, E. I., and Pabo, C. O. (2000). Combining structure-based design with phage display to create new Cys(2)His(2) zinc finger dimmers. *Structure*, **8**, 739–750.
- Zhao, Q., Chasse, S. A., Devarakonda, S., Sierk, M. L., Ahvazi, B., and Rastinejad, F. (2000). Structural basis of RXR-DNA interactions. *J. Mol. Biol.*, **296**, 509–520.

Further Reading

- Bates, A. D. and Maxwell, A. (2005). *DNA Topology*, Oxford University Press, Oxford.
- Harrison, S. C. and Aggarwal, A. K. (1990). DNA recognition by proteins with the helix-turn-helix motif. *Ann. Rev. Biochem.*, **59**, 933–969.
- Hitomi, K., Iwai, S., and Tainer, J. A. (2007). The intricate structural chemistry of base excision repair machinery: implications for DNA damage recognition, removal, and repair. *DNA Repair*, **6**, 410–428.
- Pabo, C. O. and Sauer, R. T. (1992). Transcription factors: structural families and principles of DNA recognition. *Ann. Rev. Biochem.*, **61**, 1053–1095.
- Patikoglou, G. and Burley, S. K. (1997). Eukaryotic transcription factor-DNA complexes. *Ann. Rev. Biophys.*, **26**, 289–325.
- Redinbo, J. R., Champoux, J. J., and Hol, G. J. W. (1999). Structural insights into the function of type IB topoisomerases. *Curr. Opin. Struct. Biol.*, **9**, 29–36.
- Sarai, A. and Kono, H. (2005). Protein-DNA recognition patterns and predictions. *Ann. Rev. Biophys. Biomol. Struct.*, **34**, 379–398.
- Steitz, T. A. (1990). Structural studies of protein-nucleic acid interaction: the sources of sequence-specific binding. *Quart. Rev. Biophys.*, **23**, 205–80.
- Travers, A. A. and Buckle, M., Eds. (2000). *DNA-protein Interactions, a Practical Approach*, Oxford University Press, Oxford.

Index

A

A tract, 51–53, 56, 70–73, 134, 142, 191–192, 207, 258, 268, 271
 A/T tracts, 70, 134
 A'-RNA, 61, 207
 A $\leftrightarrow$ B transition, 7
 A•G, 82–83, 85–87, 111, 212, 214, 217, 223
 Acceptor site, 233
 Accuracy and reliability, 5, 10
 Acridine carboxamide, 152, 156
 Acridines, 100, 145, 163, 164
 Actinomycin, 143, 148–149, 152
 Adenosine platform, 241
 A-DNA, 40–41, 43–45, 59–63, 89–91, 138, 207, 269, 276
 A-DNA double helices, 61
 Adriamycin, 146, 152
 Alkylation, 85, 187, 192
 Alkylation mismatches, 85–87
 Alternative conformations, 53
 AMBER, 14, 20, 24, 33
 Ametantrone, 156
 Amidoanthraquinones, 163
 Aminoglycoside, 238
 2-Aminopyridine, 99
 A-minor motif, 241
 Anomalous scattering, 8
 Anthracycline, 146, 149–150, 152, 162
 Anthramycin, 192, 193–194
 Antibacterial, 144, 237, 276
 Antibiotic drug design, 238
 Antibiotics, 143, 144, 158, 178, 234, 237–238
 Anticodon, 211, 218, 219, 223, 232–233, 238
 Antigene approach, 100

Antiparallel triplex, 95, 100
 Antitrypanocidal agents, 169
 Antitumor activity, 146, 162, 190–191, 193
 Antitumor antibiotics, 158
 Antiviral agents, 169
 Aptamer, 102, 112, 227, 238–240, 242
 Argonaute protein, 243
 A-RNA, 206–211, 213–214, 219–220, 223
 Artificial gene regulation, 100
 Atomic-resolution, 8
 A-tract(s), 53, 56, 70–73, 134, 142, 193, 207, 257, 267, 274
 A-tract bending, 142
 Averaged-sequence B-DNA, 47
 Azithromycin, 237

B

Backbone conformation, 10, 24, 31, 33–35, 40–41, 48, 55, 59–61, 63, 65, 74, 117–118, 135, 138, 146, 207, 223
 Backbone conformational angles, 10, 33, 35, 41, 48, 59, 61, 63, 65, 107
 Backbone-modified, 100
 β -Barrel recognition motif, 274
 Base morphology, 24, 51, 123
 Base pair flexibility, 24–28
 Base pairing, 7, 11, 23–25, 28, 33, 37, 40, 54, 81, 83–85, 87, 95, 101
 Base pairs, 2, 11, 23–27, 38–41, 45, 47–49, 51–53
 Base pair steps, 26, 39, 66, 134, 146

Base steps, 55, 57–58, 135, 272
 Base-pair displacement, 42
 Base-pair morphological features, 38–39
 Bases, 8–9, 12, 16, 20–21, 23–25, 27–28, 32, 35, 36
 Base-stacking interactions, 84
 Base-step parameters, 27, 47, 60
 BCR-ABL oncogene, 261
 B-DNA, 6, 14, 39–43, 45, 47, 49–51, 53, 55
 Bending, 24, 28, 49, 51–52, 55, 62–63, 70–72
 Bent DNA, 69–74, 265, 268–269, 275, 281
 Benzimidazole, 173, 176–177, 179, 183–185
 Benzo[a]pyrene, 82
 Benzo[e]pyridoindole, 162
 Berenil, 169, 174–175, 182, 193
 Bifurcated hydrogen bonds, 51, 71, 177, 181, 268
 B_I/B_{II} conformation, 83
 Bimolecular quadruplex, 105, 106, 108, 132, 165–166
 Bis-acridine, 152, 154, 156, 163, 167
 Bis-intercalation, 158–159, 161
 Bis-intercalators, 158–163
 Bizelesin, 191
 Bond length and angle geometries, 20
 Buckle (κ), 25
 Bulged RNA, 210
 B-Z junction, 67, 68

C

C form, 43, 45
 C•GC triplex, 88–89
 CA tracts, 54

- Calladine Rules, 57–58
 Cambridge accord, 24–25, 27
 Cambridge Crystallographic Database, 15
 Camptothecin, 278
 Cancer(s), 102, 143, 146, 188, 250
 Cancer chemotherapy, 143, 188
 Canonical DNA, 40, 274
 Canonical B-DNA, 41–42, 47, 49, 136, 182, 189, 191, 269, 274
 Canonical A-RNA, 207, 211, 214, 220
 Canonical fiber-diffraction A-DNA, 60
 Canonical forms, 43, 272
 CAP, 264, 265, 274
 Carbomycin A, 237
 Carboplatin, 190
 CC-1065, 192
 CEHS, 28
 Centromeric DNA, 85
 CGP40215A, 175–176, 182
 Chargaff, 23
 CHARMM, 14, 20
 Chemical cleaving, 14
 Chemical probe, 2, 14, 160
 Chemical shifts, 10
 Chimera, 16
 Chromatin, 1, 70, 187, 250, 272, 275
 Chromatin fiber, 273
 Chromomycin, 169–170
 Chromosomes, 101, 270, 272
 Circular dichroism, 61, 64, 102, 110
 Cis-DDP, 188
 Cisplatin, 188–189, 190, 191
 C-kit, 102, 111–112
 Clindamycin, 237
 C-loop, 241
 Cloverleaf structure, 218
 C-myc, 100, 102, 111, 165, C-myc promoter, 165
 Code-reading, 233
 Codon-anticodon, 234
 Codon-anticodon recognition, 232–233
 Complete ribosome structures, 235–236
 Complex ribozymes, 224–227
 Complex RNAs, 205, 210, 220
 Coralyne, 163
 Correlated flexibility, 33–35
 Coupling constant(s), 10, 31
 Crick, F.H.C., 1, 23, 39, 40
 Cro repressor, 257
 Crystal packing, 51–52, 54, 56, 59–61, 63, 71, 158, 161, 189, 191, 207, 209
 Crystal packing factors, 61, 207
 Crystalline fibers, 45
 Crystallization, 7, 60, 62, 72, 84, 143, 188, 205, 210, 234
 Crystallization of RNAs, 205
 Crystallographic unit cell, 5, 8
 CURVES, 28
 Cytotoxic drugs, 143
 Cytotoxicity, 162, 193
D
 D form, 45
 DACA, 152–153
 DAPI, 169, 181
 Daunomycin, 146, 148–149, 152, 156, 161–162
 Daunorubicin, 152
 DB289, 175
 DB921, 176, 181
 Decamer(s), 14, 47–54, 56–57, 61–63, 69–72, 84
 Defined-sequence polynucleotides, 40, 43
 Density functional theory calculations, 24
 Deoxyhexose rings, 124
 Derived stereochemical features, 9
 Diagonal loop(s), 106, 108
 Diarylfuran, 175
 Dickerson-Drew dodecamer, 47–49, 51, 56, 59, 69, 71–72, 136, 140, 170–175
 Dinucleotide, 6, 25, 40, 45, 49, 55–58, 70, 252–253
 Dinucleotide base steps, 57
 Diphenyl furan, 187
 Direct information readout, 250
 Direct methods, 8
 Direct readout, 132–133, 148, 256, 261, 263–265, 268
 Distamycin, 144, 174, 178–179, 180–183
 DNA cleavage proteins, 250
 DNA-bending, 269, 270–273, 275
 DNA enzyme, 117–120
 DNA nanostructures, 118
 DNA polymerase, 124, 252
 DNA polymorphism, 43–47
 DNA repair enzymes, 87, 264
 DNA strand breaks, 146
 DNA topoisomerase(s), 143, 168, 250, 277–278
 DNA topoisomerase II, 152
 DNA-Water Interactions, 136–141
 DOCK, 144, 205, 209
 Docking algorithms, 144
 Dodecamer(s), 47–52, 54, 56–57, 61, 63, 69–72, 84, 87
 Double-strand breaks, 143, 188
 Doxorubicin, 146, 149
 Drosophila, 187
 DSB-120, 192–193
 Duocarmycin, 191
 Duplex RNA, 60, 169, 205, 207, 216, 242
 Dynamics simulation(s), 13, 24, 31, 47, 50, 91, 112, 137, 144, 210, 215, 259
E
E. Coli ribosome, 231–232, 234
 E2, 72, 264, 265, 274
 E2 binding site, 62
 E2 protein, 55, 73, 274
 Echinomycin, 152, 158, 160–161
 Electron density, 2, 4–5, 7–9, 20, 137, 143, 167, 231, 235
 Electrostatic contributions, 13, 143
 Electrostatic interactions, 33, 58, 69, 99, 135, 251, 256
 Electrostatic potential, 13, 136, 171, 210, 261
 Ellipticine, 147, 152
 Empirical energy functions, 58
 Empirical force-field methods, 12
 Enamine, 238–239
 Energetics, 11–12, 134, 144
 Energies, 24, 144

- Engrailed, 256–259
 Ensembl, 11–12, 73, 133
 Errors and uncertainties, 7
 Erythromycin, 237
 Ethidium bromide, 146
 Eukaryotic chromosome,
 101, 269, 273
 Excisionase-DNA complex, 253
 Exit site, 233
 Expression platform, 227
- F**
- Fiber diffraction, 1, 3, 5–6, 15,
 39–47, 49, 51, 53, 56,
 60–62, 66, 74, 88–91,
 114, 137–138, 207–209
 First shell of hydration, 140
 First-shell water molecules,
 9, 47, 137
 3+1 Fold, 110
 Foot printing methods,
 15, 184
 Force fields, 13–14, 20, 92
 Fourier map, 2, 5
 Four-way branched junction, 115
 Four-way junction, 118, 157,
 167, 225
 Frameshift mutagenesis, 53
 Free-energy calculations, 144
 Free R factor (R_{free}), 5
 FREEHELIX, 28
 Free-radical cleavage, 73
 Friedrich's ataxia, 185
 Furamidine, 175, 181
- G**
- G-quartet, 101
 G•A, 81, 83–85, 87, 211, 214,
 218
 G•A mismatches, 83, 117
 GAL4, 260, 262
 GCN4 leucine zipper, 263
 Gene expression, 205, 227
 General transcription factor,
 267, 268
 Generalized Born solvation
 model, 13
 Genetic algorithm, 58
 Genomic sequences, 68, 259
 Gentamicin, 238
 Global bending, 71
 Global helical axis, 25
- Glycosidic angle(s), 10, 32–33,
 35, 41, 48–49, 54, 61, 65,
 81, 83, 85, 86, 104, 106,
 107, 108, 113, 124
 Glycosidic bond, 22, 32, 107,
 219
 G-quartets, 107, 112, 113, 164
 GROMOS, 14
 Groove-Binding Molecules,
 169–185
 Groove depth, 41, 135, 208
 Groove dimensions, 43, 106,
 208
 Grooves, 25, 40–41, 43, 45,
 55, 71–72, 106, 109, 114,
 117, 133–136, 138, 140,
 142, 156, 163, 210, 232,
 241, 250
 Groove width, 11, 40, 45, 53,
 57, 63, 69, 106, 134,
 171–172, 174, 180,
 273–275
 Group I ribozyme, 224
 G-tetrad, 101–106, 110–111,
 164, 166
 G-tracts, 103–104, 111
 G-tract sequences, 102
 Guanine quadruplexes,
 101–113
- H**
- Hairpin ribozyme, 224–225
 Hammerhead ribozyme,
 221–224
 Heavy atom derivatives, 205
 Helical dimensions, 6
 Helical DNA, 5, 6, 23, 27, 40,
 91, 147
 Helical parameters, 28, 38, 40,
 52, 59, 61, 62, 207
 Helical periodicity, 69
 Helical pitch, 38, 45
 Helical repeat, 6, 40–42,
 51, 60–61, 65, 69–70,
 209
 Helical repeating unit, 65
 Helical RNA Conformations,
 206
 Helical twist (Ω), 26, 39, 49,
 54–60, 66, 69, 90–91,
 104, 114, 124, 134,
 145–147, 191, 208–209,
 271, 273
- Helix axis, 6, 26, 28, 38,
 40, 42–44, 54, 66, 90,
 95, 120, 124, 207–209,
 273
 Helix bending, 24
 Helix–helix stacking, 220
 Hepatitis delta virus,
 224–227
 Heteronomous model, 73
 Hexads, 112
 High mobility group (HMG),
 189, 268
 High propeller twists, 51,
 53–54, 71–72
 Hin recombinase, 266
 Histone core, 270, 272, 273,
 275
 Histone H5, 274
 Histone octamer, 270–275
 Histone proteins, 266,
 270, 271
 HIV, 235
 HIV polypurine tract, 100
 HIV-1 retrovirus, 212–215
 HIV-1 Rev protein, 213, 215
 HMG boxes, 268
 HMG domain, 275
 Hoechst 33258, 169–174, 177,
 181, 187
 Holliday junction(s), 54, 58,
 114–119, 166–168
 Homo-DNA, 124
 Homeodomain, 251, 255–259,
 265
 Hoogsteen, 54, 81–83, 88–89,
 91–93, 95, 99, 101,
 160–161, 192, 216–217,
 241, 250–251
 Hoogsteen base triplets,
 216–217
 Hoogsteen hydrogen-bonding,
 81, 88–89, 160
 Host-guest, 178, 180
 Host-guest approach, 55–56
 HTH motif, 256–257, 259, 260,
 264, 266, 269, 273, 274
 HU, 270
 Human genome, 100, 102,
 132–133, 249
 Human topoisomerase I, 275,
 276
 Hybrid, 60, 63, 91, 96
 Hybrid polynucleotides, 60

- Hydration, 47, 49, 50, 55, 57, 61–62, 67, 117, 134
 Hydration of DNA, 137
 Hydrogen atoms, 7, 32, 41, 55, 135, 157
 Hydrogen-bonding recognition, 253–255
 Hydrogen bonds, 9, 23, 38, 51, 71, 84, 86–87, 89, 99
 Hydroxyl radical cleavage, 69
 Hypoxanthine, 227–229
- I**
- Idealized DNA, 6
 i-motif, 113–116, 165
 Inclination (η), 26, 40, 93, 95, 124, 208–209, 273
 Indirect readout, 135, 171, 250, 256, 265, 274
 Inosine, 45, 71, 82
 In silico library, 144
 Intact ribosome, 234
 Integration host factor (IHF), 270
 Intercalation, 100, 144–148, 151–153, 155–159, 163, 276
 Intrastrand, 83, 96, 191
 Interstrand covalent cross-links, 194
 Interstrand cross-link, 188, 190–191
 Intramolecular quadruplex, 105, 109–111, 164
 Intramolecular triplexes, 91
 Intrastrand cross-link, 188–191
 Intrinsic DNA bending, 70
 Isohelicity, 171, 172, 175–176, 177, 193
 Isomorphous replacement, 7, 8, 47, 64
 Isothermal titration calorimetry (ITC), 15
- J**
- JUMNA, 14
 Junction bends, 72
 Junction DNAs, 166
 Junction model, 70–72
- Junctions, 54, 58, 72, 114, 117, 157, 166–168, 224, 275
- K**
- Kanamycin A, 238
 Karplus relationship, 10
 Kinetoplast DNA, 70, 175
 Kink-turn, 241
 Kissing-loop, 212
 Klyne-Prelog system, 35
- L**
- Large RNAs, 7
 Lateral loops, 108, 110, 114
 Lattice interactions, 49
 Least-squares, 5–6, 9, 40, 71
 Least-squares fitting, 9
 LEF-1, 189
 Lerman, 1961, 144
 Leucine-zipper, 263
 Lexitropin(s), 180–182
 Linked-atom least-squares, 40
 Lividomycin A, 238
 Local axes, 25
 Local DNA deformability, 273
 Local DNA deformations, 271
 Local helical twist, 56–57, 69
 Locked nucleic acids, 95
 Loop(s), 11, 102, 104–111, 114, 118–119
 Loop-loop interactions, 212, 241
- M**
- Macrolide antibiotics, 237
 Macromolecular Structure Database, 15
 MAD phasing, 205
 Magnesium ions, 138, 223, 224
 Major groove, 25–27, 41–43, 47, 49, 51–52, 54, 57
 Major groove networks, 138, 140
 Major-groove interactions, 257–260
 Major-groove intercalation, 151–158
 Major-groove recognition, 136, 264
 Major-groove width(s), 60–63, 90, 208
 Maltese cross, 39
 MAR70, 149–150, 152
 MAT α 2, 259
- Melphalan, 188
 Mercury heavy-atom derivative, 8
 Meridional reflection, 38
 Messenger RNAs, 229
 Met J repressor, 265, 274
 Metal ions, 49, 138, 142–143, 157, 223
 2-Methyl-2, 4-pentanediol (MPD), 72–73
 Methylated bases, 82
 Microbial infections, 169
 Mini duplexes, 146
 Minor groove(s), 25, 41, 43, 45, 47–51, 53, 55
 Minor-groove alkylation, 192
 Minor-groove bending, 272
 Minor-groove recognition, 182, 264–271
 Minor-groove width, 49–51, 53, 57, 60–61, 90, 182, 190, 208, 275
 Mismatched DNA, 153, 156
 Mismatches, 56, 81–87, 96, 207, 241
 Mitochondrial DNA, 169
 Mitomycin C, 192–194
 Mixed sequence recognition, 96
 Mixed sugar puckers, 146
 Mobile water molecules, 55, 137
 Modified nucleosides, 219
 Molecular dynamics (MD), 9, 11, 13, 16, 24, 31, 47, 50, 74, 112, 137, 144, 176, 210, 215, 259
 Molecular mechanics, 13–14, 20, 33, 91
 Molecular modeling, 1, 11, 114
 Molecular replacement, 8, 205
 Monte Carlo simulation, 13, 73, 144
 Morphological parameters, 56, 59, 70, 191, 208
 mRNA, 23, 208, 227, 231–235, 237–238
 Multiple isomorphous replacement, 47, 64
 Multi-wavelength anomalous diffraction (MAD), 8
 Mu repressor, 266
 Mutagenesis, 53, 82, 256
 Mutagenic methylation, 82
 Mutagens, 87

N

Narrow minor groove(s), 47,
49, 53, 55, 71–72, 74,
114, 120, 182, 266
Negative roll, 57, 72
Neighbor-exclusion principle,
145–146
Neomycin B, 238
Netropsin, 144, 174, 178–181,
193, 266
Neutron diffraction, 55, 65, 138
NF- κ B, 55
Nitrogen mustards, 188
NMR methods, 10–11, 72, 95,
107, 110–111, 137, 157,
190, 193, 216
NMR “R factor”, 11
Nogalamycin, 149–152
Non-B-DNA conformation, 267
Nuclear Overhauser effect
(nOe), 10, 13
Nuclear receptors, 261
Nuclease digestion, 69
Nuclease digestion studies, 61
Nucleic Acid Database (NDB),
15–16, 20, 27, 28, 58–59,
63, 65, 73, 87, 97, 106, 137
Nucleosome(s), 6, 51, 70, 265,
270–276
Nucleosome core particle, 270

O

Octanucleotides, 60–61
Oligonucleotides, 1, 7–8, 10,
35, 51, 54, 61–62, 65, 67
Oligonucleotide synthesis, 2
Ordered water molecules, 71
Orientational disorder, 54
Oriented fibers, 6, 38
Osmium hexamine, 205
Overwinding, 43
Oxaliplatin, 190–191
Oxidative damage, 87
Oxytricha nova, 107

P

P4–P6 domain, 224–226
P53, 143, 261
Packing, 9, 34–35, 49, 51–52,
54, 56, 59–61, 63, 71,
157–158, 161, 189, 191,
207, 209, 224, 232, 272

Paracrystalline arrays, 6
Parallel loops, 105
Parallel triple helix, 89–91
Paromomycin, 234, 238–239,
Particle-mesh Ewald, 13, 92
PAZ domain, 242–243
PBD, 192, 226, 229, 261
Pentamidine, 169, 174, 175,
181, 187
Peptide bond formation, 229,
233, 234
Peptide nucleic acid (PNA), 92,
120, 123
Peptidyl transferase catalytic
activity, 234
P-form, 93, 120
Phase problem, 2, 6
Phenoxazone, 149
Phosphate conformation, 35,
49, 55, 59, 65, 83, 135
Pitch, 6, 38, 43, 45, 54, 66,
89, 120
PIWI complex, 243
Plasmid mobility, 69
Platinum lesions, 189
PNA, 92–93, 97, 120–121, 123
PNA-DNA structure, 120
PNA-DNA triplex, 93
Pneumocystis carinii, 174
Poly dA•dT, 73–74
Polyamide(s), 181–187, 261,
275–276
Polyamide-oligonucleotide
complexes, 185
Polygon arrangements, 62
Polymerase(s), 124, 191, 205,
251–252
Polymeric nucleic acids, 5
Polymorphism, 43, 47, 207
Polynucleotide fibers, 6, 54
Porphyrin, 102, 157–158,
164, 165
POT1, 102
Proflavine, 145–147, 152
Prokaryotic ribosome, 229
Promoter sequences, 102
Propeller conformations, 110
Propeller loops, 106, 110–111
Propeller twist (ω), 25, 39,
47–49, 51, 53–54, 56–57,
70–72, 74, 95, 208, 261,
268, 274

PROSIT, 29

Protein Data Bank (PDB), 4,
15–16, 53, 58–59, 65, 73,
87, 97, 106
Pseudoknot, 214, 216–217,
224–226, 241
Pseudorotation, 29, 31
Pseudorotation wheel, 29, 31
Psoralen, 100, 152, 157,
166–167
Purine bases, 20, 140, 227
Purine bulge, 213, 214
Purine-purine base pairs, 214
Purine:Purine mismatches,
82–85
Puromycin, 234
PyMOL, 16
Pyrimidine bases, 20, 23,
25, 140
Pyrrole-imidazole, 186, 275

Q

Quadruplex(s), 12, 101–113,
114, 132, 157, 164–166,
204
Quadruplex-binding proteins,
102
Quality and reliability, 9

R

R factor, 5–6, 11
RasMol, 16
Recognition code, 184, 264
Recombination, 54, 101,
114, 167
Reference frame, 27–28
Refinement, 5, 7, 9, 11, 14, 50,
53, 140–141, 191, 217,
220, 271
Refining structures, 9
Regulatory proteins, 162, 186,
250, 268
Repair, 1, 81–83, 85, 87, 167,
187, 190, 191–193,
249–250, 264
Repair proteins, 190, 191, 250
434 Repressor, 257, 258, 265
Repressor-operator complex, 254
Resolution, 2–9, 11, 20, 23, 40,
47, 50–51, 53, 55–56
Restrained molecular
dynamics, 11

- Retinoic acid, 261
 Reverse Hoogsteen pairs, 81
 Rhodium complex, 152–153, 155–156
 β -Ribbons, 264
 Ribosomal proteins, 229, 232–234
 Ribosome, 1, 6, 204, 221, 229–235, 237, 238, 241, 250
 Ribosome structure
 determination, 231
 Ribostamycin, 238
 Riboswitches, 205, 210, 227–228, 242, 243
 Ribozyme(s), 7, 204, 210, 216, 219, 221–227, 229, 234, 242
 Rise, 6, 10, 27, 38, 40, 42, 45, 54, 58–59, 208–209
 RNA, 20–35
 RNA double helix, 207–229, 211, 218
 RNA helices, 6, 60, 207, 228, 242
 RNA interference (RNAi), 205
 RNA sugars, 206
 RNA synthesis, 143
 RNA-drug complexes, 234, 248
 RNA-packaging, 212
 RNA-protein, 215, 242
 RNase P, 221
 RNA-splicing, 227
 Rnt1 protein, 216
 Robotic crystallization, 7
 Roll (ρ), 26–28, 39, 40, 51, 54–59, 70–72, 189, 191, 208, 216, 261, 271–272, 274
 Roll angle, 26–27, 55–57, 273, 275
- S**
- Screening, 7, 144, 205, 261
 Selenium, 8, 205
 Selenium modification, 8
 Self-cleaving RNAs, 221
 Self-splicing introns, 204
 Semicrystalline diffraction patterns, 40
 Sequence selectivity, 133, 169, 178–179, 187, 195, 253
 Sequence-dependency, 58–59, 274
 Sequence-dependent effects, 10, 56–57
 Sequence-dependent flexibility, 55
 Sequence-dependent structural features, 14, 24, 47, 49, 54, 56–57, 134–135, 171, 271
 Sheared base pairs, 82
 Simple intercalators, 146–147
 Simulated annealing, 9, 13
 Simulation programs, 14
 Single-crystal methods, 7–10
 Single-molecule force measurement, 15
 Single-stranded RNAs, 207
 siRNA(s), 205, 243
 SJG-131, 193
 Slide, 26, 51, 57, 63, 70, 274
 Sodium ions, 106–107, 138, 141, 142
 Solvent-accessible surface, 135, 137
 Sp1 transcription factor, 61
 Spermine, 65, 143
 Spine of hydration, 47, 49–50, 55, 134, 137–138, 140, 142, 174, 177, 210, 259
 Spiramycin, 237
 SPKK, 266
 SRY, 189, 269
 Stacking, 24, 53, 55, 57–58, 66–67, 81, 83–85, 96, 108
 Standard nomenclature, 20
 Standardized coordinate reference frame, 27
 STAT-3, 102
 Steroid hormones, 261
 Strand breaks, 143, 146, 187
 Strand cleavage, 90
 Strand-reversal loops, 104
 Streptomycin, 234, 238, 239, 240
 Structural genomics, 249
 Structure amplitude, 5
 Structure factor, 5, 9, 15
 Sugar pucker(s), 10–11, 28–31, 33, 35, 40–41, 48–49, 51
 Sugar pucker preferences, 31
 Supergroove, 275
 Superhelix, 269
 Supramolecular DNA, 168
 Surface plasmon resonance (SPR), 2, 15
 Symmetry, 5, 49–50, 139, 153, 157, 168, 233, 271
 Syn guanosine conformation, 66
 Synchrotron, 4, 8, 11, 45, 50, 137, 141, 220, 231
 Synthetic analogues, 121
 Synthetic DNA recognition molecules, 261
- T**
- T•AT triplex, 89
 T7 polymerase, 205
 TAR RNA, 235
 TATA box, 260, 267–268, 275, 276
 TATA-box-DNA complex, 190
 TBP protein, 267–269, 274
 TBP-TATA complex, 268, 270
 Telithromycin, 237
 Telomerase, 164, 214
 Telomerase RNA, 214, 216–217
 Telomeres, 101–102, 165
 Telomeric DNA, 101–102, 104, 110, 114, 164
 Telomeric DNA overhang, 110
 Telomeric quadruplex, 111, 164
 Tetrahymena intron, 224–225
 Tetraloop(s), 210, 211, 215–216, 224, 233, 241
 Tetramolecular quadruplex, 104, 132
 Tetra-N-methyl-pyridylporphyrin (TMPyP4), 102, 157–159, 164–167
 Tetranucleosome, 265, 272–273
 Tetranucleotide, 58, 223
 Tetraplexes, 102
 TFIIA, 260, 268
 TFIID, 267
 TFIIEA, 61, 186
 Therapeutic index, 143
 Thermal motion, 137
 Thiamine pyrophosphate, 228–230
 Thi-box, 228
 Three-center hydrogen-bonding, 53, 74

- Three-helix bundle, 257
 Three-way DNA junction, 168
 Three-way junctions, 241
 Thrombin, 102
 Tilt, 26, 28, 42–43, 51, 58–59, 70, 72, 271
 Time-resolved X-ray diffraction, 45
 TMPyP4, 102, 157–159, 164–166
 Topoisomerase, 143, 146, 152, 169, 174, 250, 277, 278
 Topoisomerase I, 143, 270, 275–278
 Topotecan, 270, 275–278
 TpA step, 55, 57, 72
 Transcription factor MCM1, 258
 Transcription(s), 1, 55, 60–61, 68–69, 100, 143, 169, 186–187
 Transfer RNA (tRNA), 204, 217–221, 231–235, 237–238
 Transition mutation, 85
 TRIBIZ, 177–178, 181
 Triostin, 158, 160
 Triple helices, 88–100, 216, 220
 Triplet mismatches, 96, 210
 Triplex, 132, 163
 Triplex DNA-Ligand, 163
 Triplex recognition, 99–100
 Trp repressor, 253, 256, 275
 Tsukuba Accord, 27, 28
 Tylosin, 237
- U**
 Unwinding, 81, 145–147, 158, 160, 169, 190, 250
- V**
 van der Waals radii, 9
 Vertebrate telomeric sequence, 108
 Virginiamycin S, 237
 Visualization, 15–16, 158, 209
 Visual Molecular Dynamics (VMD), 16
- W**
 Water networks, 138, 210
 Water-mediated contacts, 253–254, 259
 Watson, J.D., 1, 23, 39, 40, 118
 Watson-Crick base pair(s), 25, 27, 54, 83, 86, 98, 133, 211, 213, 217, 219, 233–234, 251–252
 Watson-Crick base pairing, 7, 33, 40, 81, 161, 207, 218, 241
 Watson-Crick hydrogen-bonding, 32, 81, 85–86, 88, 132
 Wedge model, 70–71
 Wide minor groove, 55, 60, 242
 Winged-helix motif, 254
 Wobble, 86–87, 211, 233
 Wobble pair, 211
 WP631, 161–162
- X**
 X and Y displacements, 26
 xDNA double helix, 122
 X-PLOR, 9, 11, 14
 X-ray crystallography, 1, 10, 29, 107, 142, 231, 249, 272
 X-ray diffraction, 2, 38, 45, 137
 X-ray diffraction patterns, 3, 6–7
 X-structure, 116
- Y**
 Yeast phenylalanine tRNA, 219
- Z**
 Z form, 43, 45, 67, 69, 82
 Z-DNA complex, 69
 Z-DNA double helix, 66
 Z-DNA in vivo, 69
 Z-DNA oligonucleotides, 65, 67
 Z-DNA tracts, 68
 Z_I, 65–66
 Z_{II}, 65–66
 Zif268, 260–261, 274
 Zig-zag arrangement, 64
 Zinc finger(s), 251, 259–263, 273–274
 Zinc finger design, 263
 Zinc-finger recognition, 260–264
 Zipper motifs, 252